

Proceedings of the first International Conference on the use of Mobile Informations and Communication Technology (ICT) in Africa – UMICTA 2014, 9-10 December 2014, STIAS Conference Centre, Stellenbosch: Stellenbosch University, South Africa.



UMICTA 2014 was co-organised by Stellenbosch University (South Africa) and Makerere University (Uganda), and hosted at STIAS in Stellenbosch, South Africa, and forms part of an ongoing collaboration effort between the department of E&E Engineering (through the MTN Mobile Intelligence Lab and MIH Media Lab) at Stellenbosch University, and the College of Computing and Information Sciences (through the AI lab) at Makerere University.

Labs and departments involved in organising the conference include:

- Department of Electrical & Electronic Engineering at Stellenbosch University
<http://ee.sun.ac.za>
- MTN Mobile Intelligence Lab at Stellenbosch University
<http://mtn.sun.ac.za>
- Makerere University's College of Computing and Information Sciences
<http://cis.mak.ac.ug/new/>
- Artificial Intelligence (AI) Lab at Makerere University
<http://cit.mak.ac.ug/cs/aigroup/>
- MIH Media Lab at Stellenbosch University
<http://ml.sun.ac.za>
- Computer Science department at Stellenbosch University
<http://www.cs.sun.ac.za/>

Proceedings of the first International Conference on the use of Mobile Informations and Communication Technology (ICT) in Africa – UMICTA 2014, 9-10 December 2014, STIAS Conference Centre, Stellenbosch: Stellenbosch University, South Africa.

ISBN: 978-0-7972-1533-7

Published by:

Department of Electrical & Electronic Engineering

Stellenbosch University

Private Bag X1

Matieland 7602

South Africa

Editors:

MJ (Thinus) Booysen - Stellenbosch University

Printed by:

Department of Electrical & Electronic Engineering, Stellenbosch University

© Authors and Editors – All rights reserved, 2014.

Copyright of composition and compilation of this publication belongs to the editors & Stellenbosch University. Copyright of each individual article rests with the author(s).

Note from the chairs

We would like to welcome you to the first international conference on the use of Mobile ICT in Africa (UMICTA 2014).

The conference provides a common platform for academic and industry partners to share ideas, present existing work, and guide future work. Academic work will be presented after a rigorous peer-review process that includes constructive feedback. Academic sessions will be interspersed with open industry-focussed discussion sessions, where invited industry partners will participate. Four keynote addresses will be delivered, aimed at setting the scene for, and provide focus to the discussions.

Mobile ICT has proven to be a useful means of addressing many of Africa's challenges, and continues to produce novel ways to improve the lives of the many the continent's inhabitants. We would like to use this conference to showcase some of the research that goes towards this goal. To try to make the research even more relevant and to progress it to implementation, we have invited our industry partners and have given them a platform to highlight their needs.

We would like to thank the sponsors of the event, especially the two main contributors, namely the International office at Stellenbosch University, and the College of Computing and Information Sciences.

We would also like to extend our gratitude to the following participating industry partners, without whom the presented research would not have been possible:

MTN, Trinity Telecoms, MiX Telematics, TomTom, Eskom, Diacoustic Medical Devices.

Our thanks go out to the keynote speakers, who have availed themselves to help set the scene and make the discussions relevant to the broader research and industry societies.

Finally, we would like to thank the reviewers, whose constructive feedback allowed the researchers to produce research of a high quality.

We trust that the conference will build bridges between the various stakeholders, and serve as a catalyst for continued collaboration between the academic institutions and the industry partners.

Conference chairs and organisers

MJ (Thinus) Booysen – Stellenbosch University

Richard Ssekibuule – Makerere University

John Quinn – Makerere University

Gert-Jan van Rooyen – Stellenbosch University

Brink van der Merwe – Stellenbosch University

UMICTA 2014 Programme

[UMICTA 2014 \(http://mtn.sun.ac.za/conference2014/\)](http://mtn.sun.ac.za/conference2014/)

[Taking place at STIAS \(-33.935208, 18.874021\)](tel:+2733935208)

Programme (issued 24/11/2014)

Tue 9 December (Metering, Crops, M2M, etc.)		Wed 10 December (Intelligent transport)	
07:30	Coffee	07:30	Coffee
08:00	Welcome (Organisers and Christoff Pauw)	08:00	Welcome (Organisers)
08:30	Keynote (Thomas Magedanz - FOKUS, Germany)	08:30	Keynote (Bart van Arem - TU Delft, Netherlands)
09:00	Academic:	09:00	Academic: Intelligent Transport systems
09:30	Human and crop disease management	09:30	
10:00		10:00	
10:30	Coffee	10:30	Coffee
11:00	Academic:	11:00	Academic: Intelligent Transport systems
11:30	Home automation and smart metering and M2M	11:30	
12:00		12:00	
12:30		12:30	
13:00	Lunch	13:00	Lunch
13:30	Keynote (Eben Albertyn - MTN)	13:30	Keynote by (Arnold van Zyl - Chemitz, Germany)
14:00	Industry open session - challenges and needs (MTN, Trintel, Eskom)	14:00	Industry open session - challenges and needs
14:30		14:30	(MiX Telematics, TomTom, ITS-SA, iSAHA, WC DoT, Radarvision, Discovery - TBC)
15:10	Presentation JA Quinn: UN telecoms data analysis	15:00	
15:30	Coffee	15:30	Coffee
16:00	Academic: Machine learning, e-health and gaming	16:00	Academic: ITS poster session (not peer-reviewed):
16:30		16:30	Four final year projects.

18:30 [Social function at Longridge Wine Estate \(-34.017024, 18.832990\)](tel:+2734017024)

Detailed programme on next page

M2M, smart metering, disease management, machine learning, and gaming - 9 December 2014

From	To	Authors	Title	Page
09:00	09:20	Godliver Owomugisha, John A Quinn and Ernest Mwebaze	Automated Vision-Based Diagnosis of Banana Bacterial Wilt Disease and Black Sigatoka Disease	1
09:20	09:40	Joviah Tuhaise, John A Quinn and Ernest Mwebaze	Pixel Based Classification Methods in Cassava Brown Streak Disease (CBSD) for Automatic Symptom Measurement	6
09:40	10:00	Haji Ali, Hussein Suleman and Ulrike Rivett	Developing Mobile Graphic Reminders for Reinforcing Compliance in Tuberculosis Treatment in Africa: User Evaluations	11
10:00	10:20	Martin Mubangizi, Ricardo Andrade-Pacheco, Micheal Smith, John A Quinn and Neil Lawrence	Malaria surveillance with multiple data sources using Gaussian process models	16
11:00	11:20	Douw Du Plessis and Peter Jan Randewijk	Smart Home Energy - Management System for Demand Side Mangement	21
11:20	11:40	Guy Sawyer and Marthinus Booyesen	Presentation of a Home Automation Solution with Seamless Integration and Vast Expansion Potential	28
11:40	12:00	Ronald Steinke, Asma Elmangosh, Thomas Magedanz, Joyce Mwangama, Neco Ventura and Andreea Ancuta Corici	An OpenMTC platform based interconnected European - South African M2M Test-bed for Smart City Services	35
12:00	12:20	Philip J C Nel, Marthinus Booyesen and Brink van der Merwe	ICT-enabled solutions for smart management of water supply in Africa	40
16:00	16:20	Harry Mafukidze and Riaan Wolhuter	Design and Development of a Satellite Based Water Resources Monitoring System	45
16:20	16:40	Dirk Brand and Brink van der Merwe	Comment Classification in a South African Language Domain	50
16:40	17:00	Manrich van Greunen and Herman Engelbrecht	A comparison of Quad-tree and Voronoi-based spatial partitioning for dynamic load balancing	57

Intelligent Transport Systems - 10 December 2014

From	To	Authors	Title	Page
09:00	09:20	Dominique Ter Huurne and Johann Andersen	A Quantitative Measure of Congestion in Stellenbosch using Probe Data	62
09:20	09:40	Rose Nakibuule and John Quinn	Evaluation of low cost vision-based Traffic flow monitoring in crowded cities	68
09:40	10:00	Innocent Ndibatya and Marthinus Booyesen	Modelling of inter-stop minibus taxi movements: Using machine learning and network theory	74
10:00	10:20	Nelson Ebot Eno Akpa, Marthinus Booyesen and Marion Sinclair	The impact of average speed over distance (ASOD) systems on speeding patterns along the R61	78
11:00	11:20	Marianne Vanderschuren and Nothando Khumalo	Visibility improvement through Information Provision regarding Sun Glare: A Case Study in Cape Town	83
11:20	11:40	Jarrett Engelbrecht, Marthinus Booyesen and Gert-Jan van Rooyen	Recognition of driving manoeuvres using smartphone-based inertial and GPS measurement	88
11:40	12:00	Cornelius T. Le Roux and Jacobus Engelbrecht	Automated Landing of an Intelligent Unmanned Aerial Vehicle in Crosswind Conditions using Total Energy Control	93
12:00	12:20	Claudia Struwig and Johann Andersen	Performance Measurement Trends in the implementation of Intelligent Transport Systems (ITS) within the South African (Public) Transportation Environment	98

List of reviewers

Andrew de Bruin , Stellenbosch University, South Africa
Antoine Bagula, University of the Western Cape, South Africa
Arnold Barnard, Stellenbosch University, South Africa
Arnold van Zyl, Chemnitz University, Germany
Bart van Arem, Technical University Delft, Netherlands
Brink van Der Merwe, Stellenbosch University, South Africa
David Fourie, Stellenbosch University, South Africa
Dieter Barnard, Siemens, Germany
Gerhard Lamprecht, RadarVision, South Africa
G-J van Rooyen, Stellenbosch University, South Africa
HA Engelbrecht, Stellenbosch University, South Africa
Hector Elliot, Western Cape Government, South Africa
Izak Nel, MiX Telematics, South Africa
Jacobus Engelbrecht, Stellenbosch University, South Africa
James Whidborne, Cranfield University, United Kingdom
John Quinn, Makerere University
Kees Hoogendoorn, Siemens, Germany
Koos van Zyl, iHASA, South Africa
Madri Engelbrecht, Stellenbosch University, South Africa
Marianne Vanderschuren, University of Cape Town, South Africa
Martin Weiss, Tritnel, South Africa
Mike Smith , Makerere University
MJ (Thinus) Booyesen, Stellenbosch University, South Africa
Monica Giannini, PluService, Italy
Nathalie Mitton, Inria, France
Riaan Wolhuter, Stellenbosch University, South Africa
Richard Ssekibuule, Makerere University
Ronelle Burger, Stellenbosch University, South Africa
Steve Kroon, Stellenbosch University, South Africa
Thomas Magedanz, Technical University of Berlin, Germany
Thys Cronje, Diacoustic Medical Devices, South Africa
Truhann van der Poel, Lattech, South Africa
Walter Karlen, University of British Columbia, Canada
Yusuf Kaka, MTN, South Africa

Keynote Speakers

Thomas Magedanz



Thomas Magedanz (PhD) is full professor in the electrical engineering and computer sciences faculty at the Technische Universität Berlin, Germany, leading the chair for next generation networks (www.av.tu-berlin.de) since 2004. In addition, he is director of the next generation network infrastructure competence center of the Fraunhofer Institute FOKUS (www.fokus.fraunhofer.de/go/ngni) since 2000.

Since 25 years Prof. Magedanz is working in the convergence field of fixed and mobile telecommunications, the internet and information technologies, which resulted in many international R&D and consultancy projects centered around the prototyping of advanced Service Delivery and Control Platforms for fixed and mobile Next Generation Networks for major international network operators and equipment manufacturers. In the course of his research activities he published more than 300 technical papers/articles and his OpenXXX testbed toolkits and advanced mobile broadband network testbeds are used in many R&D labs around the globe. Most famous is the FUSECO Playground (www.FUSECO-playground.org). In addition, Prof. Magedanz is senior member of the IEEE and holds guest professorships at the University of Cape Town in South Africa and Universidad de Chile in Santiago de Chile. More details can be found at http://www.av.tu-berlin.de/menue/team/prof_dr_thomas_magedanz/.

M2M/IOT: Key Enablers for efficient Smart Cities – Status Quo and Ways Forward

Looking at the big mega trends for ICT evolution, “connectivity & convergence” represents one of the most important driver for the implementation of Smart Cities and the related application domains, such as SmartEnergy, SmartHome, eHealth, eMobility, etc.

Thanks to the past convergence of telecommunications and the internet in the context of Next Generation Networks, as well as the sensorization of our environments driving the notion of the internet of things (IOT), it is clear, that the very next big challenge will be the convergence of M2M / IOT communication platforms in order to enable an efficient ICT infrastructure for realizing smarter applications within Smart Cities. In this talk we will look at the communication requirements of some Smart City application domains, and highlight the obvious similarities.

Based on this we will outline the current state of the art in international M2M standards and the available M2M products utilized by different network operators around the world.

Finally we will show, what will be the next challenges in this connectivity and convergence evolution, where for example the inherent support of M2M communications and dynamic, close to the network edge data processing in emerging software defined 5G network environments important aspects. We will also provide a snapshot on ongoing European M2M/IOT research activities.

Keynote Speakers

Eben Albertyn - The Passionate Pioneer



Eben Albertyn is the Chief Technology Officer of MTN SA.

He obtained his M.Eng degree from RAU where after he started an IT company with friends. His passion for Engineering led him to MTN in 2000 where he was part of the Radio Planning Department. Soon thereafter Eben moved onto MTN Cameroon and spent 7 years there of which the latter two and a half as the CTO. Eben then became CTO of MTN Ghana where he had the opportunity to build the Country's 3G Network and Optical Fibre Network.

Eben left MTN at the end of 2010 to join Airtel Africa as the CTO for Africa in Nairobi. He moved back to South Africa with his family to join MTN in his current position as CTO of MTN's the South African Operation. Currently focussing in transformation of the business.

My Leadership Style: A fundamental principal that I believe in is that you reach your destination with the assistance of others and not by using them, working together as a Team is key. People, who are smarter or better than me, should be given the opportunity to do that they are great at. It is essential to make them believe in themselves and to show them that you have faith in them. Today when I look back on my own life and career it is the people who believed in me that helped me to get me where I am today. I want to be that for others.

Mentoring therefore is very important. Our success depends on our willingness to mentor others and to manage the talent we have. If you believe in your people they will not disappoint you. MY style is very goal orientated and I can be very stringent when it comes to meeting deadlines. I must admit that I always believe some things could have been done better, faster, cheaper, etc. After the fact

My Family: I prefer spending as much time as possible with my family. We go cycling, swimming, watch rugby together, or visit the farm. I am also involved with matters of the church and I have been teaching Sunday School for nearly 17 years now.

Keynote Speakers

Bart van Arem



Prof Dr Bart van Arem is a full professor Transport Modelling at the Department Transport & Planning at the faculty of Civil Engineering and Geosciences at Delft University of Technology since 2009. He received his MSc (1986) and PhD (1990) in Applied Mathematics at the University of Twente, the Netherlands. He worked at the Netherlands Organization of Applied Scientific Researchh TNO in the field of Intelligent Transport Systems from 1991-2009. He was a part-time full professor in Applications of Integrated Driver Assistance at the University of Twente, the Netherlands from 2003-2009. His research and teaching responsibilities include intelligent vehicles and cooperative road vehicle systems. He is an active member of the IEEE Intelligent Transport Systems Society and various committees of the Transportation Research Board. He has worked on numerous national and international projects in Intelligent Transport Systems. His current projects include multimodal transport modelling, modelling of traffic flows with intelligent vehicles and cooperative road vehicle systems for national and international clients. He has founded the Dutch Automated Vehicle Initiative, is director of the TU Delft Transport Institute and head of the Department of Transport & Planning at the faculty of Civil Engineering and Geosciences at Delft University of Technology.

Talking traffic: wireless traffic management

The EU, and the Netherlands in particular, have invested heavily in road-side equipment for traffic management. Wireless connectivity is now enabling tailored advices and instructions for road users, enabling unprecedented possibilities for end users, by avoiding unnecessary delays and unsafe situations and by making traffic more reliable. Wireless traffic management will also strongly affect the role of traditional road-side oriented traffic management. In this presentation, we will discuss the current state of the art of wireless traffic management, focusing on dynamic route guidance, tactical driving advice, user response and the delicate balance between user equilibrium and system optimum of a traffic system.

Keynote Speakers

Arnold van Zyl



Arnold van Zyl (born April 1, 1959 in Swellendam, South Africa) is a South-African Engineer and Professor. Since April 2012, he has been Rector of Technische Universität Chemnitz. Prof. van Zyl completed his studies in chemical engineering at the University of Cape Town and obtained his PhD in engineering in 1987. In the following, he worked at the Max Planck Institute for Solid State Research in Stuttgart and, subsequently, held various leading positions in the R&D-sector of Daimler AG in Stuttgart, Ulm and Brussels. During the period between 2001 and 2007 he represented the European transportation sector in San Diego and Brussels. From 2008 to 2012 he was Professor in the Faculty of Engineering at the University of Stellenbosch and at the same time Vice-Rector for Research and Innovation.

Arnold van Zyl was awarded Honorary Professor by Tongji University, China. Furthermore, he is a member of the Academy of Science of South Africa as well as fellow of the South African Academy of Engineering. Amongst others, he is a member of the Board of Trustees of Fraunhofer Institute for Machine Tools and Forming Technology IWU and Fraunhofer Institute for Electronic Nano Systems ENAS.

His extensive publication record includes, amongst others, 69 patents.

Location-based services: opportunities for Africa

Mobile- and smartphones have become the most popular and widespread form of personal technology on the planet, with 3.6 billion unique mobile subscribers and 7.2 billion connections globally. Since 2000, the number of connections in sub Saharan Africa has grown by 44%, compared to an average of 10% for developed regions as a whole. Latest projections indicate a penetration of 75% of mobile technology in the region by 2016.

Mobile devices combine the functions of data acquisition, processing, storage and communication as well as the precise location of the user. Given this functionality mobile- and smartphones will continue to have a profound impact on all aspects of life, from communication, to providing access to services such as education, healthcare, government, navigation and financial services.

This paper will report on the potential impact of location-enabled mobile devices on the economic and social development of sub Saharan Africa. In addition, the potential barriers to penetration of mobile technology and the concomitant challenge of digital inclusion will be discussed.

Automated Vision-Based Diagnosis of Banana Bacterial Wilt Disease and Black Sigatoka Disease

Godliver Owomugisha, John A. Quinn, Ernest Mwebaze

Department of Computer Science

Makerere University

Email: [owomugisha.godliver, jqunn, emwebaze]@cis.mak.ac.ug

James Lwasa

Department of Information Systems

Makerere University

Email: lwasaj@yahoo.com

Abstract—Machine learning has been applied in agriculture in various areas including crop disease detection and image processing systems have been developed for some crops. These crops include cotton, pomegranate plant, grapes, vegetables, tomatoes, potatoes and cassava among others. However, no machine learning techniques have been used in an attempt to detect diseases in the banana plant such as banana bacterial wilt (BBW) and banana black sigatoka (BBS) that have caused a huge loss to many banana growers. The study investigated various computer vision techniques which led to the development of an algorithm that consists of four main phases. In phase one, images of banana leaves were acquired using a standard digital camera. Phase two involves use of different feature extraction techniques to obtain relevant data to be used in phase three where images are classified as either healthy or diseased. Of the seven classifiers that were used in this study, Extremely Randomized Trees performed best in identifying the diseases achieving 0.96 AUC for BBW and 0.91 for BBS. Lastly, the performance of these classifiers was evaluated based on the area under the curve (AUC) analysis and best method to automatically diagnose these banana diseases was then recommended.

I. INTRODUCTION

Banana is the fourth most grown crop in the world after wheat, rice and maize and Uganda happens to be the second largest producer of bananas after India [1]. The crop is used as a staple food source in the country however, its growth is threatened by banana bacterial wilt (BBW) disease caused by *xanthomonas campestris pv musacearum* (XCM) [2]. The wilt originated in Ethiopia and in Uganda it was reported by Tushemereirwe et al. [3] in Kayunga district in 2001. The disease has also moved into Congo, into Rwanda and Tanzania. BBW affects all types of banana and spreads very fast causing a devastating effect hence, many farmers have lost their crops and this has led to reduction of food availability and income for banana farmers. The disease is also coupled with many costs including labor for cutting down and disposing off infected plants, de-budding the male flowers and disinfecting cutting tools.

Following the outbreak of BBW epidemics, the government of the Republic of Uganda through the Ministry of Agriculture, Animal industry and Fisheries (MAAIF) in conjunction with National Agricultural Research Organization and other key stake holders constituted a national task force in December 2001, which in November 2003 formulated long-term strategy and action plan to eradicate the disease. This strategy includes a national coordinated effort of continuous monitoring of the epidemics: awareness raising and training campaigns,

empowering all stakeholders at all (district, sub county, parish and village) levels to control the disease [4].

The problem of identifying diseases in plants is a very well known one. Farmers wait for that time when the disease gets to a late stage and the symptoms are visible to realize that the crops are diseased. However, not much can be done to control the situation by that stage, hence this study aimed at early disease detection. The symptoms are visible in the leaves, male bud, fruit and stem. The disease begins with any leaf and causes them to turn yellow, brown and later they wilt. Young affected plants become stunted and may not produce any fruits. Apart from the BBW disease, Tushemereirwe et al. [3] mention other diseases that have led to the decline in banana production in many banana growing countries in the world including Uganda and these include: banana strake virus disease and banana black sigatoka (BBS). BBS blackens parts of the leaf and normally, drying starts from the edges and eventually the entire leaf is killed.

This paper is divided into five sections: starting with the introduction. Section two presents research that has been done on crop disease detection. Section three describes how different feature extraction and classification methods were applied to achieve the objectives of the study. Results of the techniques used are evaluated in section four and the last section recommends methods that worked best and future work.

II. RELATED WORK IN COMPUTER VISION FOR AGRICULTURAL DISEASE DETECTION

Computer vision systems have been used increasingly in the food and agricultural areas for quality inspection [5], [6] and evaluation purposes as they provide suitably rapid, economic, consistent and objective assessment. They have proved to be successful for the objective measurement and assessment of several agricultural products [7]. With the advantages of superior speed and accuracy, a significant number of researchers have been attracted to apply machine vision techniques in crop disease detection.

A support vector machine technique has been used for classification and identification of foliar diseases in cotton [8]. The classification process starts by finding the best feature vector for each class and then creates the final classification system from the best results obtained. To accomplish this, the following were considered: decomposition of images into

multiple channels (R, G, B, H, S, V, I3a, I3b, and GL), application of the discrete wavelet transform up to the third level, computation of the energy for each sub-band and feature vectors. This is followed by creation of the SVM classification environment, listing of the images used for training and testing and evaluation of the best feature vectors.

Al-Hiary et al. [9] proposed an automatic detection and classification of leaf diseases and the work is divided into three parts. This begins with the identification of the infected object(s) based upon K-means clustering procedure, extraction of the feature set of the infected objects using color co-occurrence methodology for texture analysis and finally detection and classification of disease type using artificial neural network (ANNs).

Aduwo et al. [10] present an automated vision-based diagnosis of cassava mosaic disease. The proposed algorithm is based on camera-phone input to provide a more efficient solution. The methodology begins with capturing leaf images with a standard digital camera. The captured image is then processed by applying various image processing techniques such as SIFT, SURF and HSV for shape feature extraction. The image is either classified as diseased or not based on other methods like a k-nearest neighbor classifier (KNN), support vector classifier (SVC) and Naive Bayes among others. A comparison on the different classifiers was done and results for the three main datasets were produced.

Others [7], [11] have demonstrated the value of image processing in inspecting and grading the quality of agricultural and food products. An automated system for the disease detection and grading in pomegranate plant was proposed in [11]. The techniques used here include color segmentation based on linear discriminant analysis, contour curvature analysis and a thinning process, which involves iterating until the stem becomes a skeleton.

The approach in [9] uses color co-occurrence methodology for texture analysis which makes it not applicable for banana leaves. However, the developed algorithm combined the features extracted in [8], [10] and this added strength to the results. In addition to the classifiers that were used, the study investigated on the behavior of other classification techniques on the dataset and recommended the best methods. This has not been done in the past for banana diseases thus making this research new.

III. METHODS AND RESULTS

The methodology aimed at detecting the BBW and BBS diseases using automated vision-based diagnosis techniques and work was divided into 4 parts: image acquisition, feature extraction, disease classification and evaluation of the classification performance.

A. Image acquisition

A Canon digital camera of 12 megapixels was used to capture both healthy and diseased images from different banana plantations in Bushenyi district (Western part of Uganda) where these diseases are common. Samples were taken from 5

sub-counties at an average of 5 diseased plants per plantation. A total of 623 image samples was used for this study and data was organized in three sets. Set one holds 360 leaves from healthy plants, set two has 220 leaves diseased with BBW and set three has 43 leaves diseased with BBS. In order to capture clear images with descriptive details, the camera was kept in both horizontal and vertical resolution of 72dpi (dots per inch). The flash mode was off since images were taken during day time with enough natural light and the process did not involve any cutting/removal of leaves off the plant. One sample image from each set is given in the figures below.



Fig. 1. Healthy leaf



Fig. 2. Leaf affected by BBW disease



Fig. 3. Leaf affected by BBS disease

B. Feature extraction and creation of feature vectors

Most of the time, the captured images may contain many objects especially in the background and working with such images leads to inaccurate/incorrect results. These images were cropped in order to obtain the leaf part only. However, cropped images then had a white background with pixel values of 255 and working with the whole image also brings inappropriate results too. To avoid this challenge a mask was applied onto the image in order to obtain the useful segment. The region with most green pixels was identified and basing on threshold value of $\text{gray} < 200$; green components of the pixel intensities are set to one and the background is set to zero. This converts an image into binary, thus indicating the segmentation of the leaf from the background. This mask was then applied onto the original image during histogram calculation as follows: the pixels with zero components were deleted (by multiplying the mask pixel values with the pixels of the original image) and only the region where the pixels are ones was considered during histogram calculation. Color histograms were extracted and transformation was from RGB to HSV, RGB to $L^*a^*b^*$. Fig.4 is a mask of Fig.1

Shape was also considered for this study and the process of calculating shape features was based on three routines namely, thresholding at different levels, extracting of connected components, calculating morphological features for each connected component. First, each image is thresholded at



Fig. 4. Mask

gray level. Connectivity openings [12] were used to calculate all the components in each thresholded image. These are called the peak components and were used to construct a max tree which is a data structure designed for morphological image processing in order to efficiently compute features or attributes of the connected components (following the same methodology as [13]). This process was done for every image and various morphological features were calculated for the connected components. Five shape attributes were therefore chosen to be more important and these include: Area of minimum enclosing rectangle, elongation, small compactness, small perimeter and Moment of Inertia.

The minimum bounding rectangle also called minimum bounding box is the smallest rectangle that contains every point in the shape. For an arbitrary shape, eccentricity is the ratio of the length L and width W of minimal bounding rectangle of the shape at some set of orientations. Elongation, Elo , is based on eccentricity [14].

$$Elo = 1 - \frac{W}{L}$$

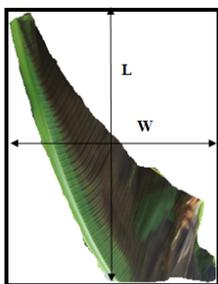


Fig. 5. Minimum bounding rectangle and corresponding parameters for elongation

The compactness measure of a shape is a numerical quantity representing the degree to which a shape is compact and one of the compact measures of shape is surface area / volume. Perimeter of an object is the distance around the outside of the object. Unlike regular shapes where at least two sides or angles are the same, irregular objects do not have these instances of symmetry and perimeter can

be determined if one takes into consideration each edge of the shape. This can either be from the left or right, bottom or top. Moment of Inertia is area (mass) times the square of perpendicular distance to the rotation axis, $I = Ad^2$.

To create feature vectors, histogram data for color components H for HSV, R for RGB and L* for L*a*b* was extracted. These components were also combined, for example HS, HV or SV and some classifiers yield better results. Another comparison was done where classification was based on the extracted shape features combined with the color histogram features. To avoid dealing with huge data and overfitting, only 50 bins were used for each case and the histograms were normalized as well.

C. Disease classification

Classifiers map an unlabeled instance of color histogram feature vectors (or a combination of color histogram feature vectors with shape features vectors) to a label. The seven classifiers used in this study were: Nearest Neighbors [15], Decision tree [16], [17], Random forest [18], [19], Extremely Randomized Trees [20], Naive Bayes [21] and support vector classifier (Linear SVM and RBF SVM) [22], [23], [24], [25]. The method used for splitting data set into training and testing was the k-fold cross-validation sometimes called rotation estimation method. The dataset was randomly split into mutually exclusive subsets (folds) of equal size of 10 [26]. The implementation platform was python with Opencv and Scikit-learn libraries. Data and source code used in achieving this are available at <https://github.com/godliver/source-code-BBW-BBS.git>.

IV. RESULTS

The choice made on which algorithm (classifier) performed best was based on the results of the AUC analysis. A comparison of the true positive rate and false positive rate for the different classifiers was done. If a classifier yields an AUC score of 1.0, then it has predicted perfectly. 0.5 is a random performance and below 0.5 means the classifier is anti-correlated with the target. Different tests were made for various color components with shape features but excellent performance was generated when the color components H and S for HSV were combined with the five shape attributes that were selected. The AUC results for the different classes (BBW, BBS and healthy) are shown in Fig 6, 7 and 8 respectively.

Of the seven classifiers, Extremely Randomized Trees yield a very high score. Both Random Forest and Extremely Randomized Trees algorithms are ensemble methods. Both algorithms are perturb-and-combine techniques specifically designed for trees. This means a diverse set of classifiers is created by introducing randomness in the classifier construction and the prediction of the ensemble is given as the averaged prediction of the individual classifiers [27]. Scikit-learn implementation combines classifiers by averaging their probabilistic prediction, instead of letting each classifier vote for a single class [18]. However, with Extremely Randomized Trees, randomness goes one step further in the way splits are computed. As in random forests, a random subset of candidate features is used, but instead of looking for the most discriminative thresholds, thresholds are drawn at random

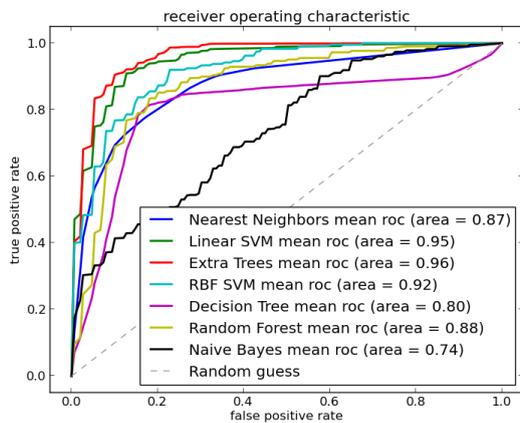


Fig. 6. HS -color components with shape attributes (AUC for BBW)

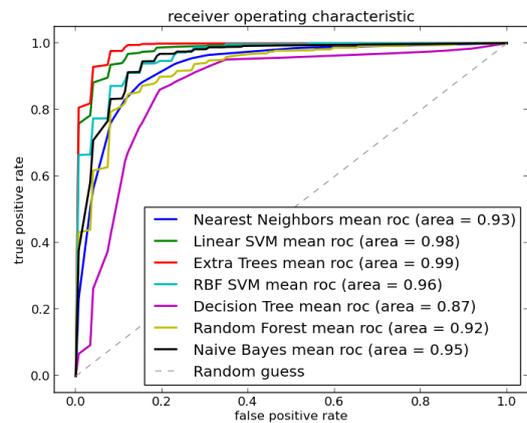


Fig. 8. HS-color components with shape attributes (AUC for healthy)

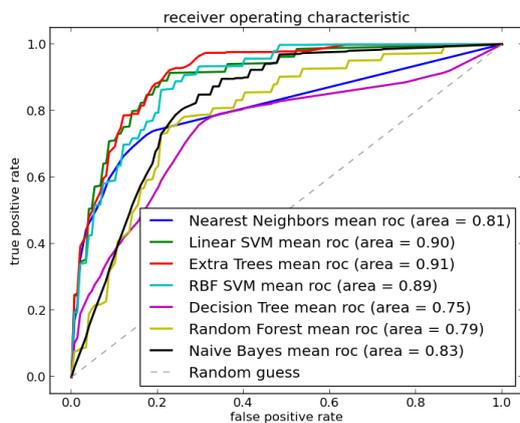


Fig. 7. HS-color components with shape attributes (AUC for BBS)

for each candidate feature and the best of these randomly-generated thresholds is picked as the splitting rule. This usually allows to reduce the variance of the model a bit more, at the expense of a slightly greater increase in bias [20]. Table 1 shows the results of Extremely Randomized Trees classifier dependent on the two leaf features. Whereas color has a greater impact than shape features, AUC performance is better when both features are combined.

V. CONCLUSION

With a very high performance of 0.96, 0.91 and 0.99 AUC for BBW, BBS and healthy classes respectively, this research has proved that there is a consistent and more accurate way to auto-detect these banana diseases rather than relying on the previous strategies that have been used in [4]. It has been shown how different feature extraction methods and classification techniques are applied systematically in the attempt to solve this problem. It is evident that the algorithm is feasible and can well identify the two diseases. Features that have been selected that work best for this application are when H and S color components are combined with the five shape features that were chosen as most important. Among the seven

	BBW	BBS	Healthy
Color	0.94	0.90	0.97
Shape	0.90	0.84	0.96
Color + Shape	0.96	0.91	0.99

TABLE I. AUC FOR EXTREMELY RANDOMIZED TREES (EXTRA TREES) CLASSIFIER

classifiers that were used, Extremely Randomized Trees is recommended because of its high performance on this data set.

The platform for automation of vision-based diagnosis of BBW and BBS diseases provides a useful direction and this work can be extended so that this works on a mobile phone device. This adds flexibility to the application since farmers are able to move with their phones to the fields and minimizes the cost of training personnel to monitor banana plants in different regions. The tool could then provide real-time information as farmers don't need to wait for experts as they can always send images to the server and then get advice. There will always be consistency of results since everyone uses the same tool. Two experts might give two different judgements on the same image, but software will always give the same answer. Other improvements that can be brought to the current work include:

- Investigating on the possibility of bananas ever getting infected by both BBW and BBS diseases.
- Adding another class to cater for healthy but mature leaves that are beginning to age or leaves affected by drought stress
- Considering features of the other parts of the plant such as the stem.

ACKNOWLEDGMENT

The authors gratefully acknowledge the AI-DEV group, department of computer science, Makerere university for the

helpful suggestions on improving this work. The authors would also like to say thanks to the team at National Agricultural Research Laboratories for being kind and helpful during the data collection stage.

REFERENCES

- [1] "The biology of bananas and plantains," *Uganda National Council for Science and Technology(UNCST) and Program for Biosafety Systems(PBS)*, 2007.
- [2] L. Turyagyenda, G. Blomme, F. Ssekiwoko, E. Karamura, S. Mpiira, and S. Eden-Green, "Rehabilitation of banana farms destroyed by xanthomonas wilt in uganda," *Journal of Applied Biosciences*, vol. 8, no. 1, p. 230–235, 2008.
- [3] W. Tushemereirwe, A. Kangire, J. Kubiriba, M. Nakyanzi, and C. Gold, "Diseases threatening banana biodiversity in uganda," *African Crop Science Journal*, vol. 12, no. 1, pp. 19–26, 2004.
- [4] O. O. W.K Tushemereirwe, D Ngambeki, "Awareness of banana bacterial wilt control in uganda: 2. community leaders perspectives," *African Crop Science Journal*, vol. 14, no. 2, p. 166, 2006.
- [5] R. H. Asankhani and H. Navid, "Qualitative sorting of potatoes by color analysis in machine vision system," *Journal of Agricultural Science*, vol. 4, no. 4, 2012.
- [6] T. Brosnan and D.-W. Sun, "Improving quality inspection of food products by computer vision: a review," *Journal of Food Engineering*, 2003.
- [7] N. V. G and H. K. S, "Quality inspection and grading of agricultural and food products by computer vision," *International Journal of Computer Applications*, vol. 2, no. 1, May 2010.
- [8] A. A. Bernardes, J. G. Rogeri, N. Marranghello, and A. S. Pereira, "Identification of foliar diseases in cotton crop," *Topics in Medical Image Processing and Computational Vision*, vol. 8, pp. pp 67–85, 2013.
- [9] H. Al-Hiary, S. Bani-Ahmad, M. Reyalat, M. Braik, and Z. Al-Rahamneh, "Fast and accurate detection and classification of plant diseases," *International Journal of Computer Applications*, vol. 17, no. 1, pp. 31–38, March 2011.
- [10] J. R. Aduwo, E. Mwebaze, and J. A. Quinn, "Automated vision-based diagnosis of cassava mosaic disease," *Proceedings of ICDM Workshop on Data Mining in Agriculture*, 2010.
- [11] S. Sanvakki, V. Rajpurohit, V. Nargund, R. Arunkumar, and P. Yallur, "A hybrid intelligent system for automated pomegranate disease detection and grading," *International Journal of Machine Intelligence*, vol. 3, no. 2, pp. 36–44, 2011.
- [12] C. Ronse, "Set-theoretical algebraic approaches to connectivity in continuous or digital spaces," *Journal of Mathematical Imaging and Vision*, 1998.
- [13] J. A. Quinn, A. Andama, I. Munabi, and F. N. Kiwanuka, "Automated blood smear analysis for mobile malaria diagnosis," *Mobile Point-of-Care Monitors and Diagnostic Device Design*, CRC Press, 2014.
- [14] Y. Mingqiang, K. Kidiyo, and R. Joseph, "A survey of shape feature extraction techniques," *Pattern Recognition*, pp. 43–90, 2008.
- [15] Z. Ma and K. Ata, "K-nearest-neighbours with a novel similarity measure for intrusion detection," *UKCI'13*, pp. 266–271, 2013.
- [16] O. Maimon and L. Rokach, "Data mining and knowledge discovery handbook, second edition," April 2010.
- [17] L. Ruey-Hsia and B. Geneva G, "Instability of decision tree classification algorithms," *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 570–575, 2002.
- [18] B. S. Leo and L. Breiman, "Random forests," *Machine Learning*, pp. 5–32, 2001.
- [19] L. Andy and W. Matthew, "Classification and regression by random forest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.
- [20] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, 2006.
- [21] H. Zhang, "The optimality of naive bayes," *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004)*, 2004.
- [22] S. R. Gunn, "Support vector machines for classification and regression," *University of Southampton, Technical Report*, 1998.
- [23] S. Keerthi, O. Chapelle, and D. DeCoste, "Building support vector machines with reduced classifier complexity," *Journal of Machine Learning Research*, vol. 7, pp. 1493–1515, 2006.
- [24] C. wei Hsu, C. chung Chang, and C. jen Lin, "A practical guide to support vector classification," *National Taiwan University, Taipei 106, Taiwan*, 2010.
- [25] N. Cristianini and J. Shawe-Taylor, "An introduction to support vector machines: And other kernel-based learning methods," *Cambridge University Press*, 2000.
- [26] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, pp. 1137–1143, 1995. [Online]. Available: <http://dl.acm.org/citation.cfm?id=1643031.1643047>
- [27] Breiman, "Arcing classifiers," *Annals of Statistics*, 1998.

Pixel Classification Methods for Automatic Symptom Measurement of Cassava Brown Streak Disease

Joviah Tuhaise

Department of Computer Science
Makerere University, Uganda
Email: tuhaise.joviah@cis.mak.ac.ug

John A. Quinn

Department of Computer Science
Makerere University, Uganda
Email: jqquinn@cis.mak.ac.ug

Ernest Mwebaze

Department of Computer Science
Makerere University, Uganda
Email: emwebaze@cis.mak.ac.ug

Abstract—The rapid geographical expansion of the Cassava Brown Streak Disease (CBSD) pandemic has devastated cassava crops in East and Central Africa. This has necessitated an upsurge of surveillance efforts for the disease. To monitor CBSD, surveyors deal with single fields of plants at a time. Diagnosis is performed by examining a cross-section cutting of the cassava root tuber. A score of severity of disease is visually assigned to the plant based on the percentage of necrotised root. This method tends to be sub-optimal since scores are highly subjective due to operator variability. This study investigates various computer vision techniques that could be employed to standardise the scoring. Our investigation follows five stages. In stage one images were acquired using mobile devices. In stage two, different techniques are employed to obtain an annotated data set that can be used to train a classifier. Stage three, several classifiers are employed to classify each pixel of the crop images as healthy or necrotised. In stage four, the performance of these classifiers is evaluated based on the Area Under the Curve (AUC), Mean Absolute Error (MAE) and R^2 score. Nearest Neighbour classifier performs best with a R^2 score of 0.789. To assess operator variability, we compare two sets of predictions from two different surveyors; a confusion matrix is used express the variability in scores assigned.

I. INTRODUCTION

Cassava (*Manihot esculenta*) is an extremely important crop in Africa, 200 million people in the continent depend on it. In Sub-Sahara Africa, cassava can represent up to 60% of the daily calorie intake, and is largely consumed locally. Cassava grows in moderately poor soils with limited labor requirements and it is drought tolerant. Thus cassava is a significant food security crop, mainly in drought-stricken areas [1][2].

The leading diseases affecting cassava in Uganda and the East African region are CBSD and Cassava Mosaic Disease (CMD). CBSD is a risk to food security, because the severity of root damage caused by the disease escalates the longer it stays in the field. CBSD, which is caused by a virus, was at first confined to coastal, low altitude areas in East Africa, but since the mid-2000s the disease has spread speedily, affecting Tanzania, Uganda, Kenya, Rwanda and Burundi. The coverage of the disease presently is around 80% of crops in Uganda and around 20% of crops in Rwanda and Burundi [3]. CBSD is a more significant cause of crop loss in these regions than was earlier believed [4] since the disease causes both quantitative and qualitative decrease in total root yield by rotting of roots thus making them unmarketable and unpalatable. For cassava plants infected with CBSD, the

major part affected is the tuber/root of the plant. To monitor CBSD, surveyors deal with single fields of plants at a time. When out in the field, normally they dig up a set of plants in selected gardens and examine five cross-section cuttings of the root. A score of severity of disease is allocated to the plant based on the average percentage of necrotised root of all five cross-sections examined. However, visual assessment of the symptoms to determine the score of severity of disease of a root by an expert may differ from the score of severity by another thus rendering this method inconsistent. To quantify the problem of operator variability, we asked two experts to assign scores of severity of disease to the same cassava cross-section cutting of the root, the results were obtained and the level of disagreement was obtained.

This paper presents an innovation to overcome this challenge. We present computer vision techniques for using camera-enabled mobile devices to automatically assign a score of severity of necrosis directly. This means survey workers with lower levels of training can be used for the surveys, and thus reducing survey costs. Given expert-annotated single images of infected cassava tubers, we also demonstrate classification of cassava root tubers based on pixel information.

II. RELATED WORK

According to Mahlein et al. [5], Spectral Vegetation Indices (SVIs) have been shown to be useful for an indirect detection of plant diseases. Nevertheless, these indices have not been evaluated to detect or to differentiate between plant diseases in crops. Their study developed specific Spectral Disease Indices (SDIs) for the detection of diseases in crops. Sugar beet plants and the three leaf diseases *Cercospora* leaf spot, sugar beet rust and powdery mildew were used as sample diseases. Hyperspectral signatures of healthy and diseased sugar beet leaves were assessed with a non-imaging spectroradiometer at different developing stages and disease severities of pathogens. To develop hyperspectral indices for the detection of sugar beet diseases the best weighted combination of a single wavelength and a normalised wavelength difference was thoroughly searched testing all potential groupings. The optimised disease indices were tested for their ability to detect and to classify healthy and diseased sugar beet leaves. With a high accuracy and sensitivity healthy sugar beet leaves and leaves, infected with *Cercospora* leaf spot, sugar beet rust and powdery mildew were classified. Spectral disease indices were also successfully applied on hyperspectral imaging data and on non-imaging data from a sugar beet field.



Fig. 1. Image sample as captured by camera

Smith and Camargo [6] developed an image-processing based algorithm to automatically identify plant disease visual symptoms. The study described an image-processing based method that identifies the visual symptoms of plant diseases by analysis of colored images. Results showed that the developed algorithm was able to identify a diseased region even when that region was represented by a wide range of intensities.

A hybrid intelligent system from color imagery for grape leaf disease detection was suggested by [7]. The system consisted of three core parts: grape leaf color segmentation, grape leaf disease segmentation and classification of diseases. The system was able to classify the image of grape leaf into three classes which were scab disease, rust disease and no disease.

Smith and Camargo [8] performed an image pattern classification for the identification of disease causing agents in plants. A machine vision system for the classification of the visual symptoms of plant diseases was implemented by analysis of colored images. A set of image features was extracted from each diseased region. Feature selection was then performed to identify which of these provided most information about the image domain. A Support Vector Machine (SVM) was used as a learning machine and cross-validation was the discrimination method used to identify the best classification model.

III. ASSESSMENT OF OPERATOR VARIABILITY

A. Methods

1) *Image Acquisition*: Image samples of cross sectional cut cassava tubers placed on a plain board were captured from Namulonge Crops resources Research Institute, Uganda. 15 root discs from the same genotype of cassava were cross sectionally cut, placed on a plain board and captured with a standard digital camera on mobile device.

2) *Expert Annotation*: Expert annotation was a process where an expert visually assigned a score of severity of disease to the sample images used.

B. Results

1) *Confusion Matrix*: To assess operator variability, by comparing two sets of predictions of two different surveyors, a confusion matrix was used to determine how these predictions differed. The results are shown in Table I:

TABLE I. SURVEYOR SCORE CONFUSION MATRIX

Scores	1	2	3	4	5
1	1	0	0	0	0
2	0	6	1	0	0
3	0	0	2	3	0
4	0	0	1	3	2
5	0	0	0	0	2



Fig. 2. Cropped image



Fig. 3. Binary image

Referring to Table I:

- The *rows* correspond to the score results as assigned by Surveyor2.
- The *columns* correspond to the score results as assigned by Surveyor1.
- The *diagonal elements* in the matrix represent the number of same score results that both surveyors assigned to the same image.
- The *off-diagonal elements* represent the score results that were assigned differently by both surveyors to the same images.
 - Off-diagonal row elements represent Surveyor1's score results that differed from Surveyor2's score results. E.g. In the second row, one image was assigned a score of 2 by Surveyor1 and Surveyor2 assigned the same image a score of 3.
 - Off-diagonal column elements represent Surveyor2's score results that differed from Surveyor1's score results. E.g. In the fourth column, three images were assigned a score of 4 by Surveyor2 and Surveyor1 assigned the same image a score of 3.

Based on the outcome of the confusion matrix, it is seen that different scores are assigned to the same image by two different surveyors. This shows how this method has problems with operator variability and therefore an automated system is a more feasible solution as compared to the surveyor visual assignment method.

IV. AUTOMATED SYMPTOM MEASUREMENT

A. Methods

1) *Image Segmentation*: In this study, images were cropped manually using a cropping tool. However, cropped images had a white background and working with the whole image also brings inaccurate results too. To eliminate the white background, only non-pure white pixels in the image were extracted automatically. Samples of the resulting cropped image and binary image are shown in Figure 2 and Figure 3 respectively.

A threshold was applied separately to each cropped root image so as to obtain its respective binary image as shown above. Coordinates of the non-pure white pixels were obtained. Using these coordinates, from the respective binary image, each pixel was labeled healthy or necrotised.

2) *Ground Truth Data*: Image pixel data was extracted from the original image and its corresponding binary image. From the original image, the *RGB* pixel data with the corresponding location coordinates, (i, j) were extracted.

3) *Classifier Training*: In the classification, the method used for splitting data set into training and testing was the k-fold cross-validation sometimes called rotation estimation, the data set was randomly split into mutually exclusive subsets of approximately equal size[9]. The classifiers used were; Nearest Neighbors[10], Decision tree[11][12], Random Forest [13][14], Naïve Bayes [15] and Support Vector Machine (Linear SVM)[16][17].

To evaluate classifier performance, four performance measures were implemented, i.e., Receiver Operating Characteristic (ROC), AUC and predictive accuracy score, which evaluated how good the classifiers performed when distinguishing necrotised pixels from healthy pixels. MAE and R^2 score, the coefficient of determination evaluated how good the classifiers performed in the overall prediction of the percentage of necrotisation in the root. The data and source code are available at <https://github.com/tjovia/CBSD.git>

B. Results

1) *Accuracy per pixel*: To choose the best performing classifier, basing on accuracy per pixel, results of AUC and predictive accuracy score with cross-validation were obtained. In this section the results for both methods will be presented. A ROC graph is a technique for visualising, organising and selecting classifiers based on their performance. Given a classifier and an instance, there are four possible outcomes, i.e., true positive (*TP*), false negative (*FN*), true negative (*TN*) and false positive (*FP*) [18].

Table II shows the results for AUC obtained for the different classifiers for the RGB color space. If a classifier yields a 1.0, then it is a perfect test, 0.9 to 0.99 is an excellent test, 0.8 to 0.89 is a good test, 0.7 to 0.79 is a fair test, 0.6 to 0.69 is a poor test, 0.5 to 0.59 is a failed test and below 0.5 the classifier is negatively correlated with the target.

With Predictive Accuracy Score, the accuracy of the test approximates how effective the algorithm is by showing the probability of the true value of the class label; summing it all up, it evaluates the overall effectiveness of the algorithm [19]. The higher the probability, the higher the predictive accuracy score. Four sample images were used to determine the probabilities as shown in Table III.

2) *Accuracy per root sample*: Performance measures used to determine best performing classifier, basing on accuracy per root sample, were Mean Absolute Error (MAE) and R^2 Score, the Coefficient of Determination. y_i , the actual percentage of necrosis, was calculated by dividing the number of necrotised pixels in an image i by the total number of pixels in the

TABLE II. AUC OF CLASSIFIERS FOR *RGB* COLOR SPACE

Sample Image	Image1	Image2	Image3	Image4
Naïve Bayes	0.96	0.96	0.96	0.96
Linear SVM	0.97	0.96	0.97	0.97
Decision Tree	0.96	0.96	0.97	0.96
Nearest Neighbors	0.96	0.96	0.96	0.96
Random Forests	0.97	0.97	0.97	0.97

TABLE III. ACCURACY SCORE OF CLASSIFIERS FOR *RGB* COLOR SPACE

Sample Image	Image1	Image2	Image3	Image4
Naïve Bayes	0.90	0.89	0.90	0.89
Linear SVM	0.92	0.91	0.91	0.92
Decision Tree	0.92	0.92	0.92	0.92
Nearest Neighbors	0.92	0.92	0.92	0.92
Random Forests	0.92	0.92	0.92	0.92

TABLE IV. MEAN ABSOLUTE ERROR FOR THE DIFFERENT CLASSIFIERS

Classifier	Mean Absolute Error
Naïve Bayes	0.049
Linear SVM	0.059
Decision Tree	0.052
Nearest Neighbors	0.049
Random Forest	0.053

image i and this was compared to \hat{y}_i , the predicted percentage of necrosis. In this section the results for both methods will be presented. The MAE was used to measure how close predictions of the overall percentage of necrotisation of a root were to the actual percentage of necrotisation of the root. The MAE estimated over N is defined as;

$$\text{MAE}(y, \hat{y}) = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|. \quad (1)$$

where \hat{y}_i is the predicted value of the i -th sample and y_i is the corresponding true value.

The results for (MAE) for the different classifiers are shown in Table IV.

The R^2 score, the coefficient of determination was calculated and this provided results on how well the overall percentage of necrotisation is predicted by the model. If \hat{y}_i is the predicted value of the i -th sample and y_i is the corresponding true value, then the score R^2 estimated over N is defined as,

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (2)$$

where $\bar{y} = \frac{1}{N} \sum_{i=1}^N y_i$.

The results for R^2 score for the different classifiers are shown in Table V.

Naïve Bayes classifier and Nearest Neighbors classifier had the lowest (MAE) of 0.049 while the other classifiers had a slightly higher MAE. This meant that Naïve Bayes classifier and Nearest Neighbors classifier were more reliable models

TABLE V. R^2 SCORE FOR THE DIFFERENT CLASSIFIERS

Classifier	R^2 score
Naïve Bayes	0.691
Linear SVM	0.605
Decision Tree	0.688
Nearest Neighbors	0.789
Random Forest	0.662

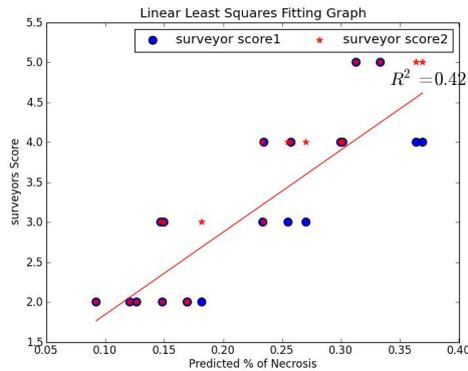


Fig. 4. Surveyor Score and Predicted necrosis Percentage

TABLE VI. R^2 SCORE FOR THE DIFFERENT RELATIONSHIPS

Relationship	R^2 score
Surveyor Score and Predicted necrosis Percentage	0.42
Actual necrosis Percentage and Predicted necrosis Percentage	0.94
Actual necrosis Percentage and Surveyor Score	0.48

compared to the other classifiers. By comparing the classifiers used, the Nearest Neighbors classifier had the highest result of the R^2 score of 0.789. This means that 79% of the total variation in Actual Percentage Necrosis is determined by the linear relationship between Nearest Neighbors Percentage Necrosis and the Actual Percentage Necrosis. Because of the highest result of the R^2 score compared to other classifiers, Nearest Neighbors classifier proved to be the more reliable the model.

The score of necrosis that was visually assigned by an expert from Namulonge Crops resources Research Institute, the predicted score of necrosis and the actual score of necrosis for the sample images was compared. The performance measure used was Linear least squares fitting and the goal was to ascertain the relationship between different pairs of these variables. And these were:

- Surveyor1 and 2 Score and Predicted necrosis Percentage.
- Actual necrosis Percentage and Predicted necrosis Percentage.
- Actual necrosis Percentage and Surveyor1 & Surveyor2 Score.

Linear least squares fitting graphs were plotted as shown in Figures 4, 5 and 6.

R^2 is a measure of how close the data are to the line and the higher the R^2 , the better the model fits the data. Based on the

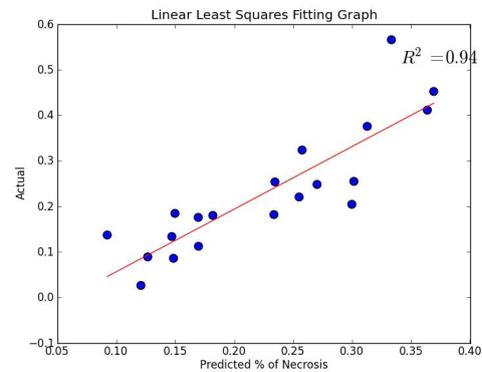


Fig. 5. Actual necrosis Percentage and Predicted necrosis Percentage

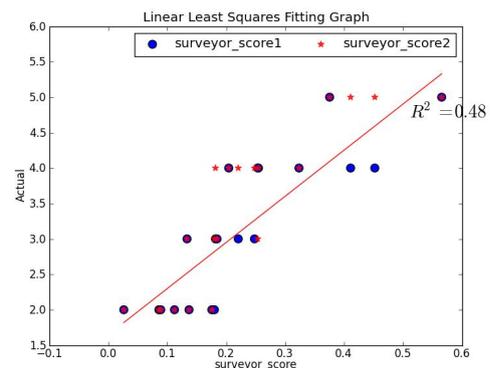


Fig. 6. Actual necrosis Percentage and Surveyor Score

results in Table VI, the R^2 score of actual necrosis percentage and predicted necrosis percentage is very high compared to actual necrosis percentage and surveyor score. This means from the results in Table VI, the model designed will predict the overall percentage of necrosis of a root well. Based on the results of the R^2 score after the linear squares fitting, it is seen how an automated system is a more feasible solution as compared to the method of surveyor visual score assignment. From the results of the R^2 score of actual necrosis percentage and predicted necrosis percentage being 0.94 and the highest as compared to actual necrosis percentage and surveyor score with an R^2 score of 0.48, an automated score assignment to the necrotised root is a more feasible solution as compared to the method of surveyor visual score assignment.

V. CONCLUSION

The research has proved that an automatic symptom measurement system specifically one that assigns a score to a necrotised cassava tuber infected with CBSD is a feasible solution, and has consistency advantages as compared to the surveyor visual score assignment method. If used, it would avert the challenge of inconsistent data collection by surveyors and would speed up the process of developing new cassava varieties that are resistant to CBSD. It has been shown how different classification techniques have been applied to automatically assign the score to the cassava tuber. Five classifiers were tested to get results; all classifiers

performed well though some performed better than others. Nearest Neighbors classifier and Naïve Bayes classifier had the lowest MAE, however Nearest Neighbors classifier had the highest result of the R^2 score and based on that, it was proved to be the more reliable the model as compared to the other four classifiers. The model was assessed and the results proved that it was a more feasible solution as compared to the method of visual assignment of the score of necrosis. In the assessment of the feasibility of the model, based on the results of the confusion matrix and the R^2 score after the linear squares fitting, it was shown how an automated system is a more feasible solution as compared to the method of surveyor visual score assignment. From the confusion matrix, the difference in scores by the surveyors to the same image showed how the surveyor visual score assignment has problems with operator variability. From the results of the R^2 score of actual necrosis percentage and predicted necrosis percentage being the highest as compared to actual necrosis percentage and surveyor score with an R^2 score, it was demonstrated that an automated score assignment to the necrotised root is a more feasible solution as compared to the method of surveyor visual score assignment.

The work can be incorporated in to a mobile version. This is planned to be done by using Open Data Kit (ODK) which is a is an open source suite of tools that enables data collection on mobile phones and data submissions to a central server. This could then improve on the monitoring of cassava brown streak disease by providing real-time information because not only experts but volunteers or farmers can take the images and then they can be uploaded on to a server at the research institute where they can be processed. Furthermore, it could improve on the prediction and optimisation of plant protection measures since there is consistency of results.

ACKNOWLEDGMENT

The authors would like to thank the AI-DEV group in the School of Computing and Informatics Technology, Makerere University for giving thoughts, advice and ideas for improvement of the study and Namulonge Crops resources Research Institute, Uganda for the support.

REFERENCES

- [1] A. L. Chávez, T. Sánchez, G. Jaramillo, J. Bedoya, J. Echeverry, E. Bolaños, H. Ceballos, and C. A. Iglesias, "Variation of quality traits in cassava roots evaluated in landraces and improved clones," *Euphytica*, vol. 143, no. 1-2, pp. 125–133, 2005.
- [2] N. Nassar and R. Ortiz, "Cassava improvement: challenges and impacts," *The Journal of Agricultural Science*, vol. 145, no. 2, pp. 163–171, 2007.
- [3] B. for Farming in Africa, "Cassava," June 2013. [Online]. Available: <http://www.b4fa.org/biosciences-and-agriculture/cassava/>
- [4] P. Ntawuruhunga and J. Legg, "New spread of cassava brown streak virus disease and its implications for the movement of cassava germplasm in the east and central african region," *USAID, Crop Crisis Control Project C3P*, 2007.
- [5] A.-K. Mahlein, T. Rumpf, P. Welke, H.-W. Dehne, L. Plmer, U. Steiner, and E.-C. Oerke, "Development of spectral indices for detecting and identifying plant diseases," *Remote Sensing of Environment*, vol. 128, no. 0, pp. 21 – 30, 2013. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0034425712003793>

- [6] J. Smith and A. Camargo, "An image-processing based algorithm to automatically identify plant disease visual symptoms," *Biosystems Engineering*, vol. 102, no. 1, pp. 9–21, 2009.
- [7] A. Meunkaewjinda, P. Kumsawat, K. Attakitmongcol, and A. Srikaew, "Grape leaf disease detection from color imagery using hybrid intelligent system," in *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2008. ECTI-CON 2008. 5th International Conference on*, vol. 1. IEEE, 2008, pp. 513–516.
- [8] A. Camargo and J. Smith, "Image pattern classification for the identification of disease causing agents in plants," *Computers and Electronics in Agriculture*, vol. 66, no. 2, pp. 121–125, 2009.
- [9] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *IJCAI*, vol. 14, no. 2, 1995, pp. 1137–1145.
- [10] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*. Morgan kaufmann, 2006.
- [11] M. N. Anyanwu and S. G. Shiva, "Comparative analysis of serial decision tree classification algorithms," *International Journal of Computer Science and Security*, vol. 3, no. 3, pp. 230–240, 2009.
- [12] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.
- [13] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [14] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.
- [15] D. M. Farid, L. Zhang, C. M. Rahman, M. Hossain, and R. Strachan, "Hybrid decision tree and naïve bayes classifiers for multi-class classification tasks," *Expert Systems with Applications*, 2013.
- [16] V. N. Vapnik, "Statistical learning theory," 1998.
- [17] P. Bartlett and J. Shawe-Taylor, "Generalization performance of support vector machines and other pattern classifiers," *Advances in kernel methods: support vector learning*, pp. 43–54, 1999.
- [18] T. Fawcett, "An introduction to ROC analysis," *Pattern recognition letters*, vol. 27, no. 8, pp. 861–874, 2006.
- [19] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation," in *AI 2006: Advances in Artificial Intelligence*. Springer, 2006, pp. 1015–1021.

Developing Mobile Graphic Reminders for Reinforcing Compliance in Tuberculosis Treatment in Africa

Haji Ali Haji
ICT4D Research School
University of Cape Town
Rondebosch, South Africa
hhaji@cs.uct.ac.za

Hussein Suleman
Department of Computer Science
University of Cape Town
Rondebosch, South Africa
hussein@cs.uct.ac.za

Ulrike Rivett
Department of Civil Engineering
University of Cape Town
Rondebosch, South Africa
ulrike.rivett@uct.ac.za

Abstract-A mobile graphic reminder is part of an application that reminds a patient about the need to follow the routine of taking medicine, and helps to monitor this process. The program is especially helpful for patients with limited literacy, language barriers or deaf. The purpose of this paper is to present and discuss (1) the benefit potential of visual-based communication in the medical context; (2) how graphics as reminder interventions to support tuberculosis (TB) treatment were designed and developed; and (3) how these graphics were evaluated. Thirty-four people, including TB patients, TB health workers and academics from the University of Cape Town, South Africa and Zanzibar, Tanzania participated in the evaluation exercise. The findings revealed that participants interpreted the meaning of most of the graphics correctly. It also found that the applications of images in the medical context might have potential to support patient treatment compared to other mobile interventions. The developed graphics are then embedded with mobile application on supporting TB patients to adhere to treatment through reminder methods. The paper contributes to mobile health (ICT4D) of developing an approach of mobile graphic-based reminder applications with literacy level, language and resource constraints.

Keywords-mobile graphic reminder; ICT4D; tuberculosis; visual communication

I. INTRODUCTION

Pictures and other visual objects such as graphs, symbols and diagrams are extremely prevalent today. People understand and remember what they see much more readily than what they hear or read [1]. Visual communication can be defined as communication through visual aids and is described as the conveyance of ideas and information in forms that can be read or looked upon [1, 2].

Several studies have been conducted regarding the use of images in medical contexts. The research conducted by Tran et al. [3] described transferring images via wireless messaging networks using camera phones to assist the diagnosis of skin diseases. The idea was that the patients could capture their infected skin areas and submit images to a doctor's phone through multimedia systems. The results found that the use of images in diagnosis provided quicker treatment and allows a physician to view and clearly understand the patient's problem rather than in text and speech.

However, the use of images as a reminder system in health contexts is a new area, particularly in Africa. Text-based and speech-based reminder systems present challenges in the developing world. The majority of African countries are faced with the problems of language barriers¹ and illiteracy. In South Africa, for example, there are 11 official languages [4] as well as high illiteracy rates. In Zanzibar, Tanzania however, all people speak one language (Kiswahili), but the problem of illiteracy exists. In these contexts, visual communication is relevant as it is largely free from language and literacy barriers.

This study aims to find out whether a graphical application is more applicable than a text or voice application in supporting tuberculosis (TB) patients in their treatment process. TB patients often forget to take their medicine as scheduled by health professional [5], which leads to difficulty in curing the disease. The number of disabilities and deaths caused by TB continue to increase. According to WHO (world health organization), approximately nine million people are infected annually in the world and almost one million die each year from TB [5]. South Africa is one of the countries with the highest burden of TB, with the WHO statistics [5] giving an estimated incidence of 500,000 cases of active TB in 2011 [6]. It is about 1% of the population. In Zanzibar, in the years 2011 and 2012, 546 and 537 (respectively) new cases of TB were recorded [7]. Zanzibar has a population of 1.3 million [8]. The most commonly found reasons [9] for missed medication are forgetting, family commitments, poor health and competing employment commitments. One way in which to avoid missing medication is by encouraging and motivating patients using reminder methods. This may lead a patient to be cured in the first phase of the TB treatment period. The minimum time for treating TB is six months [5], if a patient properly follows the treatment regulations as prescribed by health professionals. The use of mobile telephones may help patients in their treatment [5]. Mobile communication technology has been used as an intervention that reminds the patients to follow their treatment regimens.

The most commonly used mobile phone services to support patient's treatment are text messaging [10] and phone

¹ Scholars estimate that there are around 3000 spoken languages on the Africa continent [20].

call [11]. Patient received a phone call or SMS (short message service) remind him about a disease care or medication adherence. Compared to phone call system the text message has potential to work in the areas where mobile network is unpredictable, particularly in remote rural areas in developing world. The SMS also offers low-cost services than telephone call. However, both interventions require language skills. The user must be able to read and understand the language.

These technologies: text-based [9][12] and speech-based [13][14], have had use by some people to a limited extent. In this study the graphic-based reminders are proposed as an intervention to support compliance to TB treatment. Unlike text-based or voice-based, the graphic-based application is generally free from language and illiteracy barriers, which enables every person to understand the meaning of a particular reminder. This paper reports on the results of an experiment conducted to evaluate the use of graphics that would support TB treatment through a graphic reminders method.

In a preliminary study [15] that was conducted in Zanzibar from July to August 2013, the participants suggested various TB reminders to include in the proposed method. The collected texts reminders were then designed and developed into graphic reminders. The contribution of this paper is to present findings from the evaluation of graphics and illustrate how the development of a mobile application-based graphic reminder can support TB patients to adhere to their treatment.

The rest of this paper describes the motivation and advantages of the use of graphics as reminders in health contexts and how the collected text reminders were designed and developed into graphic reminders. Section 2 gives an overview of how the graphics were developed. To ensure that the graphics would be understood, the researcher conducted a survey to evaluate the graphics, as described in section 3. In section 4, the results of the survey are presented. Section 5 provides the conclusion.

II. DESIGNING GRAPHIC INTERVENTIONS

A. Development Process

The Adobe Photoshop software and Wacom tablet were used in developing the graphic reminders. The Wacom tablet was originally used to sketch prototype graphics, before Adobe Photoshop was employed to finalize the chosen graphics. Figures 1 and 2 are illustrated examples.

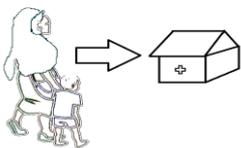


Fig. 1. Go to clinic prototype.



Fig. 2. Consultation prototype.

The advantage of the prototype sketches is that it helps to eliminate ambiguity [16]. It also helps to identify and address problems, for example missing, confusing or misunderstood features. As described in the methodology and results sections (Figures 3 to 17), after sketching several prototypes,

the researcher selected the best sketches and transferred this to Adobe Photoshop for editing, adding colour and design. The graphics were then shown to a number of people for testing. The idea behind this is that an initial design is presented to different people. They provide feedback and suggestions for improvement. These are processed by the developer, who then presents a more refined design. The people provide feedback once again. The process is repeated so that, at each stage, the sketches evolve towards the final design. The desired intention of this research was to answer the followings two questions.

- 1) How can we design, graphic reminders that can fit into various mobile phone screens?
- 2) Are the developed graphics understandable to every person; do they convey the intended meaning?

B. The Graphic Design Principles

Lawson [17] described that a good graphic designer is able to get their clients' messages across using a highly visual approach. The development of graphics in this study considered two principles of visual design as suggested by Impekable [18] - these are consistency and contrast.

- 1) *Consistency*: Consistency means creating a graphic that fits together at different resolutions and on mobile apps, and making sure that the same elements are being repeated to match each graph symbol, such as the same typeface, colour or gradient style.
- 2) *Contrast*: Contrast happens when two related elements are different. Great difference means great contrast. To make contrast work, the differences between the two graphics must be obvious. The differences are in size, colour or type.

Furthermore, mobile phone screens vary in size, contrast and type. Thus, the development considered the size and type of images in order to be compatible with different phones.

C. Graphics Integration with Voice

After the development of graphics is completed, the next part of the project will integrate those graphics with audio. The two languages to be used are English and Kiswahili. These languages are used as the project has two case studies: Zanzibar, Tanzania and Cape Town, South Africa. The choice of language considers the number of people who speak those languages in the relevant areas. In Zanzibar all people speak Kiswahili. This will enable them to understand the meaning of the voice that they will hear once a message is loaded onto the phone as a ringtone. The audio will tell the user that a reminder has been triggered to his or her phone. On the other hand, the majority of Cape Town residents understand English, so it also enables them to understand the meaning easily.

III. METHODOLOGY

A. Study Design

The study sought to find useful images to be used in an application that will support patients in their treatment process through a reminder method. The researcher conducted a survey to evaluate the developed graphics. Thirty four people participated in the evaluation. These include TB

health workers, TB patients (both inpatient and home-based-care patients), researchers and academics. Participants were from Cape Town and Zanzibar as the selected case study² locations of the study. These locations included the following participants:-

Cape Town – included participants from the University of Cape Town. This included academics, researchers and postgraduate students. A total of 17 people participated.

Zanzibar – a total of 17 people participated. These included participants from the State University of Zanzibar, which included researchers and academics, and MnaziMmoja hospital, where TB health workers and TB patients participated. MnaziMmoja hospital is the referral hospital in Zanzibar where all TB patients from central clinics are referred to.

TB health workers and TB patients are the target users of this application, but input from researchers, academics and students helped ensure that the graphics were understood by all. Furthermore, researchers and academics have expert knowledge in the graphic development context and could give suggestions on technical issues, such as graphic appearance and typeface. The participants categorised into three groups;

- Group A: Patients,
- Group B: Health workers, and
- Group C: Academics, researchers and students.

B. Ethical clearance

Ethical clearance for the study was granted by the University of Cape Town, and the Ministry of Health and Social Welfare in Zanzibar.

C. Mode of Evaluation

The evaluation was conducted between September and December 2013. The evaluation was first conducted in Cape Town then Zanzibar and then Cape Town. The two round evaluations conducted, two times in each site. All graphics were printed in colour and circulated to participants individually. The testing criterion was to ensure that each image is understood and conveys the correct meaning of a particular reminder. The reason for paper based evaluation is that once participant provided the feedback it was easy for the researcher to improve image using pencil together with participant, before modified through software.

Figure 3, for example, represents the reminder that the patient is to go to a clinic. Therefore, during the evaluation, it was observed if the participant could describe that correct meaning of the particular graphic. The feedback given helped to improve the development of graphics. Ten graphics were evaluated in the first round as shown: Figures 3 to 12.

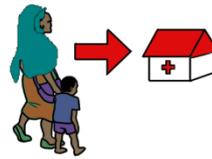


Fig. 3. Patient reminded to go to the clinic.



Fig. 4. Patient reminded to consult with a doctor.



Fig. 5. Patient reminded to take medication.



Fig. 6. Patient reminded to submit his or her smear sputum to doctor for checking.



Fig. 7. Patient reminded to take a glass of milk.



Fig. 8. Patient reminded to eat a healthy meal.



Fig. 9. Patient reminded not to cough in this manner.



Fig. 10. Patient reminded to cough in this manner.



Fig. 11. Patient reminded to collect medication for upcoming days.



Fig. 12. Patient reminded to visit a clinic when feeling unwell.

IV. FINDINGS

This section analyzes the results of the survey that was conducted to evaluate the graphics.

A. Participants

Table 1 shows the details of respondents who participated in the experiments. Of the respondents, 53% (n=18) were male and 47% (n=16) were female. One-third of participants were patients (n=11). All patient participants were from Zanzibar. Four of them were inpatients and seven were out-patients (home-based-care patient).

² According to [21], Case study is an empirical investigation about a contemporary event that exist in its real life context.

TABLE I. CHARACTERISTICS OF PARTICIPANTS

Gender	User Group				Total	
	Patient	Health Worker	Academic/ Researcher	Student	No. Part	%
Male	5	1	10	2	18	52.9
Female	6	3	6	1	16	47.1
Total	11	4	16	3	34	100

No. Part=Number of Participants, %=Percentage of Respondents

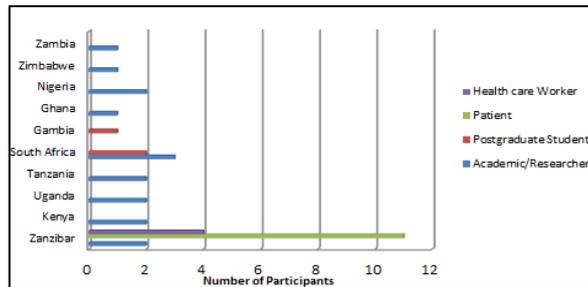


Fig. 13. The Academic and Hospital participants

Figure 13 illustrates the details of where participants come from. Those participants from the University of Cape Town were found they are from different nationalities. As shown in Figure 13 the original countries of participants were: Kenya, Uganda, Tanzania, South Africa, Gambia, Ghana, Nigeria, Zimbabwe, Zambia, and Zanzibar. Different participants were included, to ensure that the graphics conveyed the same intended meaning and did not confuse people regardless of first language, cultural affiliation etc.

B. Results

The findings exposed that participants interpreted the meaning of most of the graphic correctly. There were some graphics that were not clearly understood by some. All respondents interpreted figures 3, 4, 9, 10, 11 and 12 correctly. However, the majority of participants understood the content of figures 5, 6, 7 and 8, but these were found to be more confusing especially for participants from group A and C.

Most of the respondents suggested adding a glass of water in Figure 5. They indicated that showing a pill only does not give a clear message that this patient is reminded to take the medication. Instead, they suggested adding a glass of water as shown in Figure 14. Another improvement was made in Figure 6 that reminds the patient to collect their smear sputum and submit it to the doctor. Before the improvement, sputum bottles were shown in both hands of the patient and the doctor, as shown in Figure 6. The participants suggested that the bottle from the doctor's hand should be removed. This means that the reminder now is clearly interpreted as the patient submitting the sputum bottle to doctor as shown in Figure 15.



Fig. 14. Medication reminder- this figure was obtained based on the feedback in Figure 5.



Fig. 15. Submit sputum reminder- this Figure was obtained based on the feedback in Figure 6.

Furthermore, participants suggested adding a cow to the image in Figure 7. The glass of white liquid does not give a clear message. If the cow image is added in the graphic, it will clearly represent that a patient is reminded to take a glass of milk, as shown in Figure 16. Another improvement was in the meal suggestion reminder as shown in Figure 17. This graphic was obtained after the improvement of Figure 8. The majority of participants from group A and C, including illiterate patients, suggested that Figure 8 was not clear. Based on their cultural values, people do not eat vegetables with a spoon.



Fig. 16. Take milk reminder- this Figure was obtained based on the feedback in Figure 7.



Fig. 17. Get vegetable reminder- this Figure was obtained based on the feedback in Figure 8.

Four graphics out of the ten (Figures 5, 6, 7 and 8) were re-developed based on the respondents' feedback obtained during the evaluation. The other six graphics (Figures 3, 4, 9, 10, 11 and 12) were not re-developed. Based on the participants' feedback, there was no need for new graphics, though there were particular suggestions observed, such as colour, margin and image size. After the modifications have been made, the second round evaluation was conducted, in this time all images were interpreted correctly by all groups. However, again, there were some colour and margin suggestions proposed by minority of participants from group C. In addition, the majority of the participants who participated in the first round evaluation they were also involved in the second round. The second round testing also involved new participants from each group include, three participants from group A, one from group B and six from group C, and together they reached 34 participants as shown in Table 1 and Figure 13. All proposed suggestions were considered and will be adopted during the development of mobile health applications.

Furthermore, participants were asked about the use of visual-based, text-based and voice-based applications and which intervention could be more applicable to support TB patients to adhere to treatment through reminder methods. However, the participants were only shown the visual-based. They were asked about their experiences with the use of SMSs and telephone calls. It found that all participants had use SMS and phone call services before. Accordingly, when they compared these services, they advocate visual that could be more feasible than text and speech. Therefore, the study findings suggested, as found in the literatures, that the use of visual communication could be more applicable and provide a clearer understanding by everyone compared to other mobile interventions.

To summarize, this study was conducted with the aim of finding out whether the developed graphics are understandable and conveyed the correct meaning of a particular reminder and whether graphic communication could be more applicable compared with text and speech. The findings indicated that all participants described the proper meaning of every graphic. However, further suggestions were proposed to improve some graphics. The study also found that the use of graphics in the medical context might have potential of providing a clearer understanding of the intended meaning compared to text and speech.

Additionally, there are extra observations obtained, such as people's cultural values and religious perceptions on what the graphics look like in term of dress and eating style. The findings of this study contribute to the broader project of the development of a mobile reminder application that would be used to support TB treatment adherence in Africa.

The mobile reminder prototype will be developed on Android platform. The motivation to use Android is that according to Michael [19] by the end of 2013 Android was at 78.4% of the operating system market share, making it one of the most used mobile operating systems globally. This leads Android phones to become cheaper and cheaper every day. At present, the price of Android phones is closer to the price of feature phones, which are widely available in Africa continent. However, a next study will find out to what extent the target population use Android phones.

After developing a mobile reminder application another study will be conducted to measure the performance of visual applications compared with text or speech. The prototype application will be installed into different Android phone models with different screen sizes.

V. CONCLUSION

The use of visual objects in communication is very powerful and an invaluable resource in medical contexts, particularly for nonliterate people. However, we cannot draw conclusion based on this paper findings. It found that the use of graphics in the medical reminder system may give patients a better understanding of the intended meaning of the message compared with text and speech. It was also suggested by every participant that the use of graphics could be more applicable in supporting TB treatment. The developed graphics shown in this paper then are embedded with a mobile reminder application to support TB patients in their treatment process. The reminder system helps patients to follow the routine of taking medication.

ACKNOWLEDGEMENT

This study is supported by the Hasso Plattner Institute. We also would like to express sincere appreciation to everyone who has participated in the experiment.

REFERENCES

- [1] S. Armstrong, *Information Literacy: Navigating and Evaluating Today's Media*, Shell Education, 2008.
- [2] S.K. Card, D. Jock, and S. Ben, *Readings in Information Visualization: Using Vision to Think*, Morgan Kaufmann, 1999.
- [3] K. Tran, M. Ayad, J. Weinberg, A. Cherng, M. Chowdhury, S. Monir, and C. Kovarik, "Mobile Teledermatology in the Developing World: Implications of a Feasibility Study on 30 Egyptian Patients with Common Skin Diseases," *Journal of the American Academy of Dermatology*, 64(2), 302-309, 2011.
- [4] B.P. Tshotsho, "Mother Tongue Debate and Language Policy in South Africa," *International Journal of Humanities and Social Science*, 39-44, 2013.
- [5] S.W. Moreno, "Global Tuberculosis Report 2012," Geneva, Switzerland, Incidence and Risk Factors for Tuberculosis in HIV-Guyatt, 2012.
- [6] TB Facts, "Tuberculosis Statistics for South Africa," <http://www.tbfacts.org/tb-statistics-south-africa.html>, 2011.
- [7] H.A. Haji, R.M. Ali, and K.H.A. Suleiman, "Opportunities in the Establishment of Mobile Healthcare System for HIV and TB Patients in Zanzibar," *International Journal of Information and Communication Technology Research*, Vol. 3 No. 9, pp. 296-300, 2013.
- [8] NBS, "National Bureau of Statistics," Available <http://www.nbs.go.tz/>, 2012.
- [9] K. Akhter, S. Dockray, and D. Simmons, "Exploring factors influencing non-attendance at the diabetes clinic and service improvement strategies from patients' perspectives," *Practical Diabetes*, 29(3), 113-116, 2012.
- [10] Zurovac D, Talisuna AO, Snow RW, "Mobile Phone Text Messaging: Tool for Malaria Control in Africa," *PLoS Med* 9(2): e1001176. Doi:10.1371/journal.pmed.1001176, 2012.
- [11] Pai, N., Supe, P., Kore, S., Nandanwar, Y. S., Hegde, A., Cutrell, E., and Thies, W., "Using automated voice calls to improve adherence to iron supplements during pregnancy: a pilot study," In *Proceedings of the Sixth International Conference on Information and Communication Technologies and Development: Full Papers-Volume 1* (pp. 153-163). ACM, 2013.
- [12] E. Barclay, "Text messages could hasten tuberculosis drug compliance," *The Lancet*, 373(9657), 15-16, 2009.
- [13] A. Parikh, K. Gupta, A.C. Wilson, K. Fields, N. Cosgrove, and J. Kostis, "The effectiveness of outpatient appointment reminder systems in reducing no-show rates," *The American journal of medicine*, 123(6), 542-548, 2010.
- [14] D. Hanauer, K. Wentzell, and N. Laffel, "Computerized Automated Reminder Diabetes System (CARDS): E-mail and SMS cell phone text messaging reminders to support diabetes management," *Diabetes technology & therapeutics*, 11(2), 99-106, 2009.
- [15] H.A. Haji, H. Suleman, and U. Rivett, "Mobile Graphic based Communication: Investigating Reminder Notifications to Support Tuberculosis Treatment in Africa," In *Health Information Science*, pp. 204-211, Springer Publication, 2014.
- [16] B. Berenback, D. Paulish, J. Kazmeier, and A. Rudorfer, *Software and Systems Requirements Engineering, In Practice*. New York: McGraw-Hill, 2009.
- [17] B. Lawson, "How Designers Think: The Design process Demystified," *Routledge*, 2006.
- [18] Impekable, "The Principles of Good Visual Designer," <http://impekable.com/the-principles-of-good-visual-design/>, 2013.
- [19] Michael Oleaga, "iOS vs. Android vs. Windows Phone Market Share 2013," *Google Smartphones OS Hits 78 Percent Globally As Apple Inc. Drops Despite Strong iPhone Sales*, 2013.
- [20] G. Barbara, *Africa: Teacher Created Resources*, 6421 Industry Way, Westminster, CA, USA, 1999.
- [21] R.K. Yin, *Case study research: Design and methods*, Vol. 5, Sage, 2009.

Malaria surveillance with multiple data sources using Gaussian process models

Martin Mubangizi*, Ricardo Andrade-Pacheco[†], Michael Smith*, John A. Quinn*[‡] and Neil Lawrence[†]

*Makerere University, Kampala, Uganda
 {mmubangizi,msmith,jquinn}@cit.ac.ug

[†]University of Sheffield, UK
 {acq11ra,N.Lawrence}@sheffield.ac.uk

[‡]UN Global Pulse, Kampala, Uganda

Abstract—A statistical framework for monitoring the health of a population should ideally be able to combine data from a wide variety of sources, such as remote sensing, telecoms, and official health records, in a principled manner. Gaussian process regression is commonly used to visualise disease incidence by interpolating values across a map; in this article, we show how it can be extended to deal with many different types of information by introducing a flexible covariance structure across data sources. Combining many data sources in a single model provides a number of practical advantages, such as the ability to automatically determine the importance of each data source through likelihood optimisation, and to deal with missing values. We show the basic idea with an application of malaria density modeling across Uganda using administrative records and remote sensing vegetation index data, and then go on to describe further extensions such as the incorporation of human mobility data extracted from mobile phone call detail records (CDRs).

I. INTRODUCTION

Malaria remains endemic across much of the world, in spite of mitigation measures by both governments and international agencies. Health department intervention is now principally response-driven; at those times and locations with the greatest malaria infection rates the provision of treatment needs to be able to match the number of cases without stock-outs or staff-shortages. Hence planning stock and staff deployment depends on accurate and timely information regarding the distribution of malaria cases. In Uganda, the Ministry of Health receives weekly counts of reported malaria cases from all districts. However, this data is compromised by cases of non-reporting at both the district and health center levels [1], the cases reported are often based on unverified diagnoses, and there are various other sources of measurement error.

In order to resolve ambiguity about how the disease burden is distributed, models can be constructed which relate infection levels across time and space, or incorporate covariates which provide extra information. These covariates may be environmental (rainfall levels, temperature, vegetation strength) or social (population density, migration/movement patterns, demographics), for example. In this regard, NDVI index, which is widely used to estimate vegetation density [2], turns out to be good proxy for rainfall [3] and has proved useful in identifying suitable habitats for mosquito breeding [4].

Any attempt to use remote sensing data, such as NDVI, for carrying out inference on administrative records, will face the problem of trying to mix two data sources with differing

space and time resolutions. For example, while HMIS data is reported weekly and aggregated at a district level[†], NDVI is provided a much higher resolution in a grid and is reported every 5 days.

Gaussian process regression is commonly used in epidemiology to interpolate disease counts across space. In this paper, we explain how it can be extended to a coregionalised form in order to incorporate information from covariates. By specifying a covariance structure relating a number of inputs and outputs, it is possible to combine several different types of data in a single, principled framework. We illustrate this model using weekly counts of malaria incidence by district in Uganda, and show that for certain regions, the incorporation of environmental remote sensing data can significantly improve the estimates of the infection rate compared to baseline models. We then describe how social data can be incorporated, in particular information about movements of the population derived from mobile phone call detail records.

This paper is organised as follows. Section II discusses some of the related work; Section III presents the data used and introduces the model framework. Application of the model to environmental covariates is discussed in IV, and Section V discusses use of mobility data. We conclude, with suggestions for future work, in Section VI.

II. RELATED WORK

The need to use data from multiple sources to enhance disease modeling has been an active research area [5], [6]. [7] cites challenges that this research has been faced with. This also led to search for new data sources that may provide signals of changes in disease rates, including absenteeism [8], sales of over-the-counter health products [9], emergency call centers [10], and automatic malaria diagnosis results [11]. Examples of research that has focused on using multiple data sources, such as [6], acknowledge the need for data from multiple sources in biosurveillance. BioPHusion [5], for instance, is a framework that can use real time data from several sources for awareness and timely response.

It is widely understood that determining the geographical distribution of a disease is vital in its control [12] and in estimating the cost of that control [13]. To this end, considerable effort has gone into producing risk maps of diseases at different

[†]HMIS data might be available at smaller aggregation levels, however the information available to the authors had a district aggregation.

spatial scales—by country [14], continent [15] and at global [12], [16] scale. These example risk-maps are over a long time-scale however, looking at seasonal averages. The methods we propose offer predictions of disease counts at a weekly time-scale, allowing more detailed and precise estimates for operational use. One feature these studies have in common is the use of remote sensing for disease prediction; we use this same idea but at a much shorter time-scale.

The use of mobile phone CDRs for modelling the effects of human mobility on the distribution of malaria infection is also gaining traction in the literature (for example see [17] for a review, but also see [18] for issues around this data).

Gaussian process regression in epidemiology

Gaussian Process Regression, or Kriging models, were first introduced in the 1960s for geostatistics, and since then the method however has had application across many disciplines. In brief, the method works by estimating the correlation structure of the data (over time, space or other dimension of interest) then using these estimates of correlation the values of the output can be estimated from training data. This basic regression can be extended to combine multiple output variables, by estimating their coregionalised correlations. Further, the uncertainty and absence of data can be incorporated, allowing our confidence in each data point to be taken into account. Finally the output also includes confidence intervals, giving important information about the reliability of each of the model's estimates. The use of Kriging in public health datasets is commonly used to interpolate disease incidence across space. For example, Kleinschmidt et al. [19] applied this method to malaria mapping.

III. DATA AND METHODS

A. Data

a) Uganda Health Management Information System (HMIS): HMIS, hosted at the Ministry of Health in Uganda, manages countrywide reported cases of diseases of public health importance including malaria. HMIS receives weekly counts of reported malaria cases from health centers aggregated at district level. For each week, the number of cases in each district is reported, with an associated statistic on the proportion of health centres which were included. Often the number reporting is low, causing degradation in the quality of the data, to the point where, unprocessed, the data is of little use [1].

b) Population estimates: This data was obtained from World Pop^e, which provides estimates of the number of people living in 100m square grid cells across the entire country.

c) Normalised Differenced Vegetation Index (NDVI): NDVI gives a measure of how vigorous the vegetation is across space. In this study we use vegetation index data from eMODIS obtained from the Famine Early Warning System of the United States Geological Survey (USGS FEWS)^d. This data was obtained at a spatial resolution of 250m, every five days. To reduce the computational complexity and make the

remote sensing data more representative, a population-density-weighted average of the remote sensing data was calculated for each district.

B. Methods

A Gaussian process (GP) regression is a machine learning algorithm for relating an output \mathbf{y} (e.g. disease incidence) with a set of inputs \mathbf{X} (e.g. longitude and latitude). The core assumption of this mathematical model is that there is an unobserved or latent variable \mathbf{f} that depends on \mathbf{X} , but for which we only have access through its distorted version \mathbf{y} . This unobserved variable is a Gaussian process with some mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ which depend on the inputs [20]. The distortion is given by independent random noise at each observation.

It is possible to extend GP regression to deal with many outputs, rather than just one [21]. Broadly speaking, this approach consist of defining a multiple output kernel functions able to incorporate information from different outputs and use it to model the correlation between them. Here we are interested in showing through an application how these kind of models can be used for integrating different sources of information for malaria modelling.

Assume we have d sets of outputs and inputs $\{\mathbf{y}_1, \mathbf{X}_1\}, \dots, \{\mathbf{y}_d, \mathbf{X}_d\}$, where all \mathbf{X}_j belong to the same domain. The number of observations in each set can be different and the domains of the outputs do not have to be the same. A first approach for learning all these tasks could be to model each with a separate GP. However if we know that the outputs might be correlated we could also try to model them together. This way, information from one domain can constrain the values in another.

The mathematical formulation for such coregionalised models is broadly the same as for standard, single-output GP regression. We use the same pairing of outputs and inputs from the d original sets,

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_p \end{pmatrix} \quad \text{and} \quad \mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_p \end{pmatrix}.$$

The covariance matrix is defined in a block structure, where each block contains the weighted cross-correlation. Thus, given a kernel matrix \mathbf{K} and a matrix of weights \mathbf{B} , the multiple output kernel \mathbf{M}_K is defined as

$$\begin{aligned} \mathbf{M}_K &= \mathbf{B} \otimes \mathbf{K}(\mathbf{X}, \mathbf{X}) \\ &= \begin{pmatrix} B_{1,1} \cdot \mathbf{K}(\mathbf{X}_1, \mathbf{X}_1) & \dots & B_{1,d} \cdot \mathbf{K}(\mathbf{X}_1, \mathbf{X}_d) \\ \vdots & \ddots & \vdots \\ B_{d,1} \cdot \mathbf{K}(\mathbf{X}_d, \mathbf{X}_1) & \dots & B_{d,d} \cdot \mathbf{K}(\mathbf{X}_d, \mathbf{X}_d) \end{pmatrix}. \end{aligned} \quad (1)$$

Complex covariance structures can be defined by constructing \mathbf{K} from other kernels [22] or by using a linear combination of multiple output kernels, thus defining

$$\mathbf{M}_K = \sum_{r=1}^R \mathbf{B}_r \otimes \mathbf{K}_r(\mathbf{X}, \mathbf{X}). \quad (2)$$

^e<http://www.worldpop.org.uk/data>

^d<http://earlywarning.usgs.gov/fews/>

IV. APPLICATION

A. Temporal Modeling

As mentioned earlier, malaria models can be improved by considering covariates such as NDVI index. Here we show an example for modelling both variables across time. For this task, vector autoregressive models or a general linear model might be considered as a first option for studying the relation between this two variables. However these models require the input and output variables to be sampled at regular and equal time and space intervals. This is usually resolved by the interpolation of one of the variables. One of the advantages of coregionalised GP regression is this step is not required, and the uncertainty in an interpolated value is already incorporated into the result.

For each district, we trained single GP regression model for malaria incidence and a joint model with NDVI. Then we predicted malaria incidence 180 days ahead. In the first model, the prediction only depended on past observation of the same variable. In the second model the prediction was aided by the training observations of NDVI which overlapped the period of malaria prediction.

Figure 1 shows a comparison of a single GP regression model of malaria incidence with a joint model with NDVI information. It can be clearly seen that HMIS and NDVI are strongly correlated (values are standardized in the figure), and that the joint model performs better at predicting values from the test set.

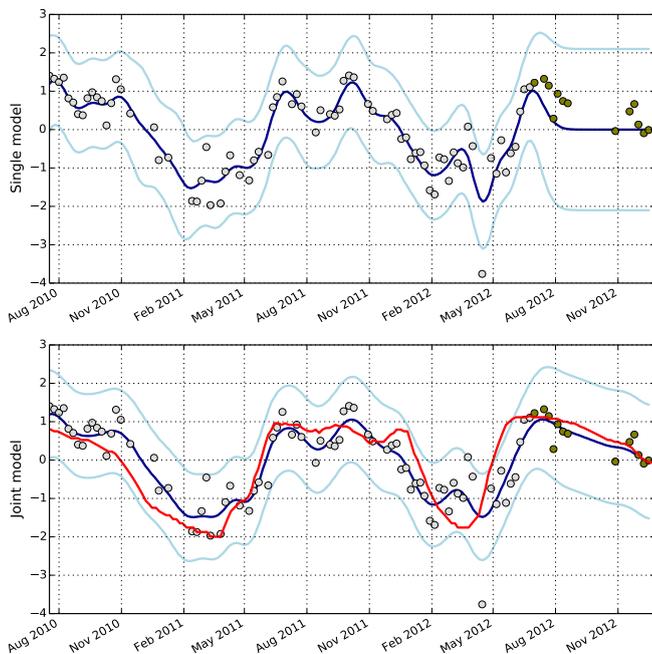


Fig. 1. Malaria incidence in Napak. The image above shows a the predictions using an independent model. The image below shows the predictions using a joint model. Training and test points are shown in gray and green circles. Predictive mean and confidence intervals are shown in solid blue lines. NDVI data is shown in red.

The similarity between malaria incidence and NDVI does not generalise across all districts. To identify those districts where there seems to be a stronger relation between these two

variables we used the the quantity

$$\beta = \frac{B_{1,2}}{\sqrt{B_{1,1}B_{2,2}}}, \quad (3)$$

where $B_{i,j}$ are the entries of \mathbf{B} (the coregionalisation matrix). Despite the similarity in the equation (3) with the definition of correlatin between two variables, it is worth highlighting that we are not giving $\beta_{1,2}$ the same interpretation.

We found that the mean squared errors (MSE) of the coregionalised model tend to be smaller than the ones from the single model, in districts with larger values of β . Figure 2 shows the ratio of MSE between the models for districts with $\beta > 0$.

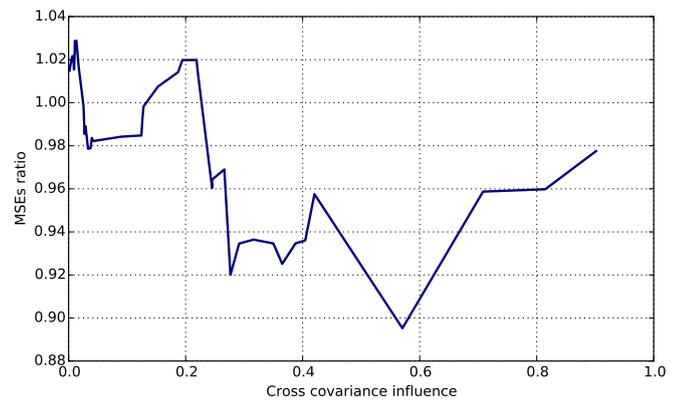


Fig. 2. MSEs vs cross covariance influence. The line shows the total MSEs of all districts with a value equal or larger than β (cross covariance influence).

Intuitively, the joint model should be as good as the single model, as in the worst scenario, where no correlation is found, \mathbf{B} would be the identity and therefore we would be assuming independence. There are however a few reasons why this intuition is not totally right and, as shown in figure 1, where we can expect the joint model to have a poor performance.

First of all, model that uses a kernel like the one in (1), but where \mathbf{B} is the identity, is not entirely independent. Although correlation across outputs is zero, by learning the parameters of \mathbf{K} with information of both outputs we are forcing the model to share information. If both variables are different, models where the kernel parameters are learnt separately can be better. By using a kernel defined as in 2 we can create a covariance structure where a covariance structure of the joint model, leads to actual independent individual covariance structures, where the kernel parameters are not shared.

Another case where the joint model can perform poorly is when outputs are not correlated, but still there are spurious correlations. In such situation, we would only be learning and sharing noise across outputs. This can be originated by the fact that two variables behave similar in for some period or because one of the variables has scarce observations that almost all learning depends on the observations of the other variable. In our experiments, we found that in district Kole $\beta \approx 1$, while there seems to be no relation between this two variables.

A computational problem that may arise due the number of elements in \mathbf{B} increases quadratically with respect to the

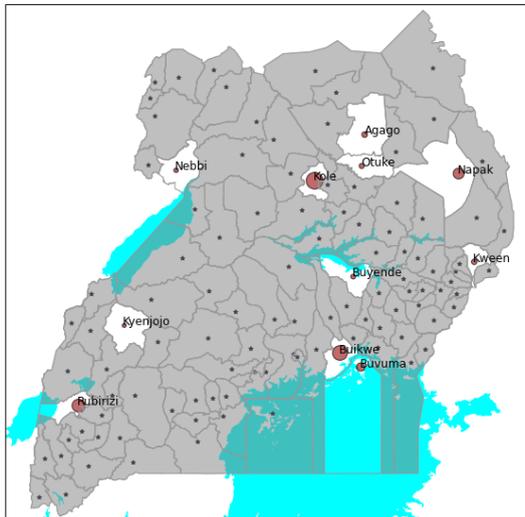


Fig. 3. Map of Ugandan districts, those with grey background have $\beta \leq 0.3$, while those with white background have $\beta > 0.3$. Points in each district have been drawn at the weighted centroid that used population as a weight.

number of outputs. Because GPs are an inference method that relies on gradient optimization, when number of parameters is large, the contribution of each one to the objective function can become negligible for some initializations.

B. Spatial Modeling

The GP coregionalised model can also help in combining HMIS data with satellite environment covariates (such as topology, NDVI, land surface temperature, land cover and land Use data) to produce a continuous surface of malaria disease risk. Here HMIS and the covariates can be treated as outputs of the model. The inputs of HMIS we can associate with reporting locations if known, but if not available then a population-weighted centroid (such as in Figure 4) can be calculated and used. The inputs of the environmental covariates will be the locations at which their values will be sampled, in this case also population can be used if its distribution is known. Since the model can exploit correlation across different outputs in space, the HMIS values will be smoothed out to generate a risk surface.

V. TELECOMS-DERIVED HUMAN MOBILITY DATA

Early models of epidemiology considered disease dispersion to depend on geographical proximity of places, or on simple gravity models of human movement, although movement patterns can be complex and significantly affect the distribution of infectious diseases [23]. Population mobility can be obtained from CDRs for example by simply counting, for each time frame, how many people moved between each pair of cell towers on a telecoms network. Thus when a single user makes or receives a call routed through cell tower i , then later makes or receives a call routed through cell tower j , we increment the count of $i \rightarrow j$ movements. This results in a transition matrix T_{CDR} , whose entries denote the fraction of people moving from one location to another.

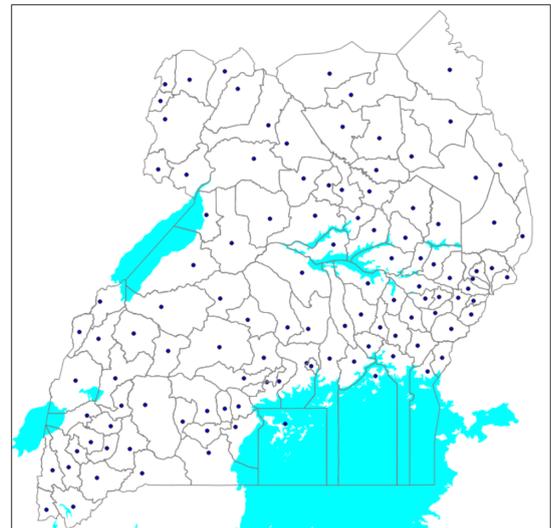


Fig. 4. Calculated positions of population-weighted district centroids.

This information, originally recorded by each tower, can be aggregated at different areas (e.g. districts), to show the average movement between them. This gives an idea of the proximity between regions that can be more informative than the actual distance between regions for analyzing infectious diseases.

Since malaria is not transmitted directly between humans, but need a mosquito as intermediary, the measure of proximity we need goes beyond human mobility. For each individual that travels from one region to another, we also need to incorporate information about infection rate in the region of origin and probability of infection the destiny region. As proxies of both quantities we can use the parasite rate and reproductive number from [12]. Finally, we should weight the movements across regions by the population in each one. Thus we can think of a transition matrix, whose elements are defined as

$$T(i, j) = T_{CDR}(i, j) \cdot P(i) \cdot RC(j) \cdot PR(i) . \quad (4)$$

Due to the lack of up to date census information in Uganda, we use values calculated from the population estimation of 2010 worldpop[§].

With this transition matrix, we can modify the distances between regions given by their centroids to include mobility information. Figure 5 shows how we shift the centroid of each district towards those that have more access to it, based on daily movements. We can use the coordinates of this new space as the inputs in a GP and then apply the methods we have discussed so far.

VI. CONCLUSION

We have presented coregionalised Gaussian process regression as a method that can support both spatial and temporal modeling of disease incidence using a range of different types of data. Whereas standard GP regression, or Kriging, is a well-established method in epidemiology in general and malaria surveillance in particular, this model provides the ability to

[§]<http://www.worldpop.org.uk/data>

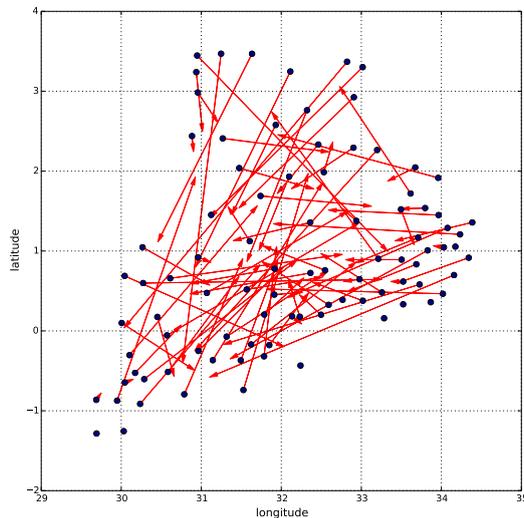


Fig. 5. Space generated by telecom data. The points correspond to the district centroids. The arrows show how each centroid is offset towards those that are closer to it in terms of the telecom data.

augment the basic regression with other data types that might be informative. In turn, while the coregionalised method has been used successfully in a number of other domains where fusion of data of different types is necessary, it has not previously been proposed in epidemiology. Using data from Uganda, we have illustrated the operation of this type of model and the types of inference it can support with remote sensing and telecoms data. We are currently collecting a more extensive dataset in order to evaluate the predictive power of these models against alternative models.

ACKNOWLEDGMENT

Author MM was partly funded by Google. Author RAP was funded by CONACYT and SEP scholarships. We are grateful to Orange Uganda for providing mobility data.

REFERENCES

- [1] World Health Organization, "Assessment of health facility data quality. Data quality report card Uganda, 2010-2011," WHO Press, Geneva, Tech. Rep., 2011.
- [2] A. R. Huete, H. Q. Liu, K. Batchily, and W. J. D. A. van Leeuwen, "A comparison of vegetation indices over a global set of TM images for EOS-MODIS," *Remote sensing of environment*, vol. 59, no. 3, pp. 440–451, 1997.
- [3] A. C. A. Clements, H. L. Reid, G. C. Kelly, and S. I. Hay, "Further shrinking the malaria map: how can geospatial science help to achieve malaria elimination?" *The Lancet infectious diseases*, vol. 13, no. 8, pp. 709–718, 2013.
- [4] S. Hay, J. Omumbo, M. Craig, and R. Snow, "Earth observation, geographic information systems and *i*₀ plasmodium falciparum/*i*₀ malaria in sub-saharan africa," *Advances in Parasitology*, vol. 47, pp. 173–215, 2000.
- [5] H. Rolka, J. OConnor, and D. Walker, "Public health information fusion for situation awareness," in *Biosurveillance and Biosecurity*, ser. Lecture Notes in Computer Science, D. Zeng, H. Chen, H. Rolka, and B. Lober, Eds. Springer Berlin Heidelberg, 2008, vol. 5354, pp. 1–9. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-89746-0_1

- [6] A. W. Moore, B. Anderson, K. Das, and W.-K. Wong, "Combining multiple signals for biosurveillance," in *Handbook of Biosurveillance*, 1st ed., M. M. Wagner, A. W. Moore, and R. M. Aryel, Eds. Elsevier Inc, 2006, ch. 15, pp. 321–331.
- [7] D. D. Angelis, A. M. Presanis, P. J. Birrell, G. S. Tomba, and T. House, "Four key challenges in infectious disease modelling using data from multiple sources," *Epidemics*, no. 0, pp. –, 2014. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S175543651400053X>
- [8] L. Lenert, J. Johnson, D. Kirsh, and R. M. Aryel, "Absenteeism," in *Handbook of Biosurveillance*, 1st ed., M. M. Wagner, A. W. Moore, and R. M. Aryel, Eds. Elsevier Inc, 2006, ch. 24, pp. 361–368.
- [9] W. R. Hogan and M. M. Wagner, "Sales of over-the-counter healthcare products," in *Handbook of Biosurveillance*, 1st ed., M. M. Wagner, A. W. Moore, and R. M. Aryel, Eds. Elsevier Inc, 2006, ch. 22, pp. 321–331.
- [10] R. M. Aryel and M. M. Wagner, "Emergency call centers," in *Handbook of Biosurveillance*, 1st ed., M. M. Wagner, A. W. Moore, and R. M. Aryel, Eds. Elsevier Inc, 2006, ch. 25, pp. 369–374.
- [11] M. Mubangizi, C. Ikae, A. Spiliopoulou, and J. A. Quinn, "Coupling spatiotemporal disease modeling with diagnosis," *Twenty-Sixth AAAI Conference*, 2012.
- [12] P. Gething, A. Patil, D. Smith, C. Guerra, I. Elyazar, G. Johnston, A. Tatem, and S. Hay, "A new world malaria map: Plasmodium falciparum endemicity in 2010," *Malaria Journal*, vol. 10, no. 1, p. 378, 2011. [Online]. Available: <http://www.malariajournal.com/content/10/1/378>
- [13] J. Omumbo, A. Noor, I. Fall, and R. Snow, "How well are malaria maps used to design and finance malaria control in africa?" *PLoS ONE*, vol. 8, no. 1, 2013.
- [14] A.-S. Stensgaard, P. Vounatsou, A. W. Onapa, P. E. Simonsen, E. M. Pedersen, C. Rahbek, and T. K. Kristensen, "Bayesian geostatistical modelling of malaria and lymphatic filariasis infections in uganda: predictors of risk and geographical patterns of co-endemicity," *Malaria journal*, vol. 10, p. 298, 2011. [Online]. Available: <http://europepmc.org/articles/PMC3216645>
- [15] H. G. M. Zour, S. Wanji, M. Noma, U. V. Amazigo, P. J. Diggle, A. H. Tekle, and J. H. F. Remme, "The geographic distribution of loa loa in africa: Results of large-scale implementation of the rapid assessment procedure for loiasis (raploa)," *PLoS Negl Trop Dis*, vol. 5, no. 6, p. e1210, 06 2011.
- [16] S. I. Hay, C. A. Guerra, P. W. Gething, A. P. Patil, A. J. Tatem, A. M. Noor, C. W. Kabaria, B. H. Manh, I. R. F. Elyazar, S. Brooker, D. L. Smith, R. A. Moyeed, and R. W. Snow, "A world malaria map: Plasmodium falciparum endemicity in 2007," *PLoS Med*, vol. 6, no. 3, p. e1000048, 03 2009.
- [17] D. K. Pindolia, A. J. Garcia, A. Wesolowski, D. L. Smith, C. O. Buckee, A. M. Noor, R. W. Snow, and A. J. Tatem, "Human movement data for malaria control and elimination strategic planning," *Malar J*, vol. 11, no. 1, p. 205, 2012.
- [18] A. Wesolowski, N. Eagle, A. M. Noor, R. W. Snow, and C. O. Buckee, "The impact of biases in mobile phone ownership on estimates of human mobility," *Journal of The Royal Society Interface*, vol. 10, no. 81, p. 20120986, 2013.
- [19] Kleinschmidt, I and Bagayoko, M and Clarke, GPY and Craig, M and Le Sueur, D, "A spatial statistical approach to malaria mapping," *International Journal of Epidemiology*, vol. 29, no. 2, pp. 355–361, 2000.
- [20] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, Cambridge, MA, 2006.
- [21] M. Álvarez, L. Rosasco, and N. D. Lawrence, "Kernels for vector-valued functions: A review," *Foundations and Trends in Machine Learning*, vol. 4, no. 3, pp. 195–266, 2012.
- [22] A. Berlinet and C. Thomas-Agnan, *Reproducing kernel Hilbert spaces in probability and statistics*. Springer, 2004, vol. 3.
- [23] A. Wesolowski, N. Eagle, A. J. Tatem, D. L. Smith, A. M. Noor, R. W. Snow, and C. Buckee, "Quantifying the impact of human mobility on malaria," *Science*, vol. 338, 10 2012.

Smart Home

Energy Management System for Demand Side Mangement

D. du Plessis

Electrical and Electronic Engineering
Stellenbosch University
Stellenbosch, South Africa
Email: douwddp@gmail.com

Dr. P.J. Randewijk

Electrical and Electronic Engineering
Stellenbosch University
Stellenbosch, South Africa
Email: pjrandew@sun.ac.za

With South Africa's energy grid currently under severe pressure due to a lack of planning, electricity prices have escalated. This has forced consumers to look toward finding innovative ways to reduce their consumption.

The aim of this research project was to create a system that assists in the reduction of household power consumption by simplifying the process of power saving. This is done by enabling consumers to remotely monitor and control their household devices.

To achieve this goal, a power management system was designed that consists of a Beaglebone black, the main electronic controller, and a circuit for measurement and control. The measurement circuit measures the power consumption, it is then sent to Trintel's SMART Platform enables users to remotely monitor and control the devices by using a web-based interactive graphical user interface.

I. INTRODUCTION

In the last decade South Africa's energy demand has significantly increased. This trend is not just confined to just South Africa but is a global one. As we entered the technological era, the amount of devices that require electricity increased immensely and this has put a big strain on the electric utility, in South Africa's case Eskom. To get ahead of this growing demand Eskom is trying to increase its capacity by building new power plants as well as operating their existing plants uninterruptedly. Another strategy which has been adopted is to put old "mothballed" plants back into operation at a very high cost. As a result the cost of electricity has increased and will continue to do so if the growth in power usage is not regulated.

Because of the continual increase in the demand for electricity and the utility's relatively static capacity to generate it cannot be reached. The relatively static capacity is mainly due to the lengthy lead time that it takes to increase infrastructure as well as the lack of capital. Fortunately there is now a constant drive towards a more sustainable society that seeks to reduce the per capita energy need. Utilities can take advantage of this intent by incentivizing the reduction in energy use and in doing so create spare capacity and buy time before new plants come online.

The main influencing contributors to energy usage levels in an average household are its geyser, stove, heating appliances, pool pumps and lights. In order to get South Africans to reduce

their energy usage Eskom have launched numerous campaigns such as the 49m [1] initiative to create awareness and inform the public on how to reduce energy consumption. The impact of these campaigns, however, has not been very successful as it depends solely on the user's willingness to change established habits and routines and make lifestyle adjustments.

The concept of this project came from Eskom's Integrated Demand Management Program (IDM). IDM was created to ensure the security of electricity supply by optimizing energy use and balancing electricity supply and demand [2]. During times of high demand the energy grid can become unstable and deviate from the operating frequency of 50Hz. In order to avoid this, Eskom runs backup jet turbine generators at a very high cost to maintain system integrity. This project applies the IDM program's concept to the general public by encouraging consumers to use less electricity by incentivizing every kWh of electricity they save, affording users the opportunity to conveniently control their household energy consumption.

Two factors that are important for a system that will help manage the energy consumption of households is its ease of use and non-intrusiveness, so as to increase the willingness of consumers to participate. Curbing electricity usage will save the user of this system money, which will double as an incentive for consumer participation as well as having the desired effect of decreasing electricity demand.

II. HISTORY OF SOUTH AFRICA'S ENERGY SECTOR

South Africa's electricity needs are dominated by Eskom, a state-owned public utility, which supplies approximately 95% of South Africa's electricity [3]. Its total generating capacity comprises a total net output of about 41GW [3]. When the fully democratic government came into power in 1994 they embarked on an ambitious programme known as Reconstruction and Development Programme or RDP. One of the goals of this programme was to provide electricity to all homes. Between 1994 and 2000 around 1.75 million extra homes were connected to the grid [4]. The 1998 White Paper on the Energy Policy of South Africa stated that the "growth in electricity demand was to exceed the generation capacity by approximately 2007" and that "long capacity-expansion lead times require strategies to be put into place" [5]. Unfortunately the new government also attempted the privatisation of the

electricity sector in the late '90s and as a result it denied Eskom's request to build new power stations in 1998. No further base-load stations were built and from early in 2006 the grid was regularly crippled due to a lack of generating capacity that led to widespread blackouts due to load-shedding [6]. The president of South Africa finally confessed in 2008, "When Eskom said to the government: 'We think we must invest more in terms of electricity generation'... We said not now, later. We were wrong. Eskom was right. We were wrong." [7]

Notwithstanding the president's confession, the government and Eskom was severely criticised for not being able to prevent the energy crisis. As a result Eskom embarked on an extensive Demand Side Management programme to reduce energy consumption. This included the distribution of more than 30 million CFLs between 2007 and 2010 [9] and funding the installation of geyser blankets to reduce heat loss, covering nearly 180 000 hot-water cylinders during 2006 in the Western Cape alone [10]. Incentives were also offered to consumers to replace stoves and geysers with gas and solar-powered versions [11]. By 2010 these interventions realised total demand savings of 2 372 MW during evening peak time, when compared to 2003 consumption [12]. By the end of 2013 Eskom's GM for integrated demand management (IDM), Andrew Etzinger, claimed that "the interventions had made a material difference in the country's power situation. With verified savings of about 3 600 MW since inception, the IDM programmes have established capacity equivalent to that of an average power station." [13]. Unfortunately the end is not yet in sight with Eskom stating as recently as November 2013 that "South Africa's power supply is stable but remains tight" [14]. This situation is set to continue until at least 2017 when two new coal fired power stations, Medupi and Kusile, each with a capacity of 4 800 MW, is expected to be completed [15].

In recent years Eskom also had to endure increased criticism due to its reliance on non-renewable energy sources such as coal. South Africa is a sunny country that offers excellent opportunities for solar power. According to the Department of Energy, South Africa's annual 24-hour global solar radiation average is about 220 W/m² versus about 100 W/m² for Europe [16].

South Africa is in desperate need for economic growth to fight the high prevalence of unemployment, some arguing for growth as high as 8% [19]. To accomplish this, high growth in energy generation capacity will be needed. An immediate net growth in the available power can be realised if consumers further cut back on their consumption. Two of the most important ways to realise this are by using more efficient appliances and employing technology to spread the power demand more evenly. A Smart energy management system will definitely help to realise this goal.

III. SYSTEM DESIGN OVERVIEW

A. Complete System Overview

The system consists of four main parts, namely the Smart Platform, modem, Beaglebone Black and the Power Measurement Circuit (PMC) as illustrated in Fig. 1. The objective of this system is to enable a user to remotely monitor a household's power consumption and then switch different zones/appliances within the household on or off in order to reduce power consumption. The Beaglebone Black is the heart of the system as it plays the role of the Main Electronic Controller and handles all the communication between the different parts.

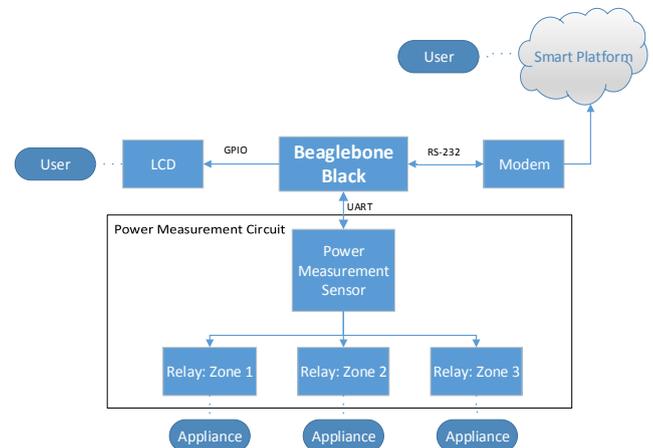


Fig. 1: Complete System Diagram

The functioning of the system can be described as follows: The power measurement sensor measures the supply voltage and current drawn by the connected load. The Beaglebone Black in turn requests a reading of the power drawn by the load from the sensor every 10 seconds. The Beaglebone Black then calculates the energy consumption per hour as this gives a much clearer understanding of how much energy is being consumed rather than just instantaneous power usage. This data is then processed by the Beaglebone Black and sent to the Smart Platform using the Sierra Wireless modem.

The user can view the data on the Smart Platform and depending on how high the power consumption is, can make an informed choice on what actions to take. Two zones/appliances can be controlled remotely from the platform by clicking on a button that sends a command back to the BBB which triggers a relay that switches off the zone/appliance or device that is connected to it.

B. Communication System Overview

Fig. 2 shows how the communication between the BBB, modem and power sensor is achieved.

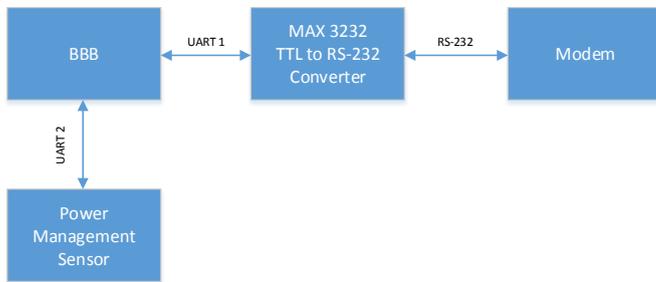


Fig. 2. Communication System

Communication between the BBB and the Modem and between the BBB and Power Sensor uses the UART protocol. Communication to the power sensor is relatively simple as both the devices have UART functionality and a direct connection could be made. The modem on the other hand communicates using the RS-232 standard. As the BBB operates at a 3.3V TTL level, a converter was needed to ensure reliable and safe communication.

C. Power System Overview

Fig. 3 shows the components and layout of the power sensor.

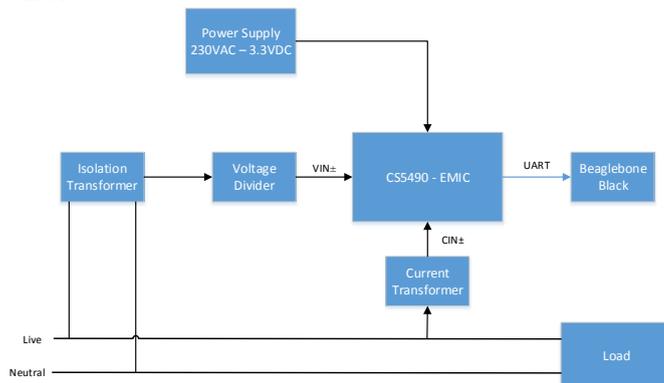


Fig. 3. Power System

The heart of the power sensor consists of the CS5490 EMIC (Energy Measurement Integrated Circuit). As the CS5490 needs a 3.3V supply voltage, a converter was designed to convert the 230VAC to 3.3VDC. For the CS5490 to accurately calculate the power consumption of a device it needs a reference voltage and reference current supplied to and drawn by the load. An isolation transformer and Voltage divider was used to supply the reference voltage and a current transformer was used to supply the reference current.

IV. PRINTED CIRCUIT BOARD DESIGN

Two Printed Circuit Boards (PCB) were designed using Altium Designer, namely a cape for the Beaglebone Black and a power measurement circuit.

A. Beaglebone Black Cape Design

To communicate between the Beaglebone Black and the modem the 3.3V TTL logic levels of the Beaglebone need to be converted to RS-232 levels. For this purpose a cape was designed for the Beaglebone to simplify connections and improve reliability. The cape connects to the Beaglebone Black with two 46 pin headers and the modem connects to the cape via a RS-232 connector. The main components incorporated in the cape are as follows:

- MAX3232 Logic Level Shifter
- LCD screen
- RS232 connector

As mentioned earlier the BBB communicates via UART and this method of serial communication is also known as TTL serial communication. This type of communication transmits one bit at a time at a certain data rate (in this case 9600bps). TTL levels will always remain between the limits of the microcontroller, namely 0V and VCC, and in the case of the Beaglebone these limits are 0V and 3.3V where 0V represents a logical 0 and 3.3V a logical 1. RS-232 serial communication works the same as TTL serial signals as it also sends one bit at a time at a certain data rate. The difference lies at hardware level. The RS-232 standard has a logical high at a negative voltage between -3 to -25V and a logical low at a positive voltage between +3V to +25V. The reason for these higher voltages on the RS-232 signal is to make it less susceptible to noise and also allows it to travel longer physical distances than the TTL signal [20].

In order to achieve this, a MAX3232 IC was used on the cape. This IC converts the RS-232 signals to TTL and vice versa while also protecting the transmitter outputs and receiver inputs against voltage spikes as high as 15kV.

The LCD screen was included on the Beaglebone cape initially to simplify debugging and coding but was later used to show real-time energy consumption measurements as well as voltage and current readings. This enables the user to get most of the crucial information on site without needing to connect to the internet.

B. Power Measurement Circuit Design

In order to make installation easier the power measurement circuit was designed that would be able to power the system, take measurements as well as to switch the loads connected to it on or off. The main components incorporated on the power measurement circuit are as follows:

- AC to DC converter
- Voltage Regulator
- Energy Measurement IC
- Isolation Transformer
- Optocouplers

To power the system a 3.3V and 5V DC power supply were needed to power the Energy Measurement IC and the Beaglebone Black respectively. An AC to DC converter was used in conjunction with a voltage regulator to convert the 220VAC from the mains to DC voltages for the system.

The Energy Measurement IC (EMIC) was chosen as it calculates power consumption using only a reference voltage and current. The power consumption data can then be read from the IC's internal registers by means of UART communication [21].

As mentioned above, the EMIC needs a reference voltage to calculate the power consumption. This can be acquired by using a voltage divider circuit. This technique is effective but can be dangerous for the user and the EMIC as high voltages are connected to the voltage divider circuit. Should something go wrong it would break the EMIC. To prevent this from happening and to isolate the circuit from the high voltage AC lines an isolation transformer was incorporated into the design.

Further protection circuitry came in the form of optocouplers. These components were used to obtain electrical isolation between the Beaglebone Black and the Power Measurement circuit. This isolation helps to protect the Beaglebone Black from voltage spikes and grounding faults as the device is especially vulnerable to external voltages when it is powered down.

V. SMART PLATFORM

To enable users to remotely communicate with and control the device a cloud platform service was setup. The Sierra Wireless Airvantage platform is a cloud platform service that allows service providers to build wireless Machine-to-Machine (M2M) applications [22]. This platform is used by Trinity Telecomms (Pty) Ltd to power their Trinity SMART platform and enable users to remotely monitor telemetry devices.

The setup of the Smart platform consists of two parts, the asset model setup and the Metric model setup. The asset model setup is used to create the asset and define its characteristics by defining its variables, events, alarms and commands. The AirVantage Configuration tool is used for this setup.

Once the asset model was created, the metric model setup could be completed. This allows the conversion of raw data coming from the asset to user understandable information that is displayed on the platform.

The Trinity SMART platform can be used to create a graphical user interface called a dashboard. This dashboard can be designed by the user to display data from telemetry systems on various graphs and meters called gadgets. The dashboard can also be equipped with buttons to send commands or data to the asset (telemetry system) [23].

The dashboard that was designed for this project can be seen in Fig. 4. Final Dashboard Layout

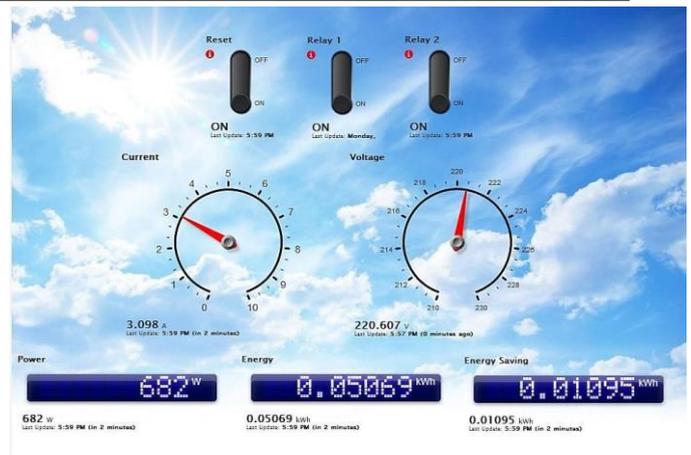


Fig. 4. Final Dashboard Layout

This dashboard receives five sets of data, namely voltage, current, power, energy consumption and energy savings. The power and energy measurements are displayed in a digital format as it gives the user the most accurate readings while the voltage and current are displayed on gauge scales to appeal to the general public by making the information more visually stimulating to the consumer. The dashboard also consists of three toggle switches, namely Reset, Relay 1 and Relay 2. These switches are used to send commands back to the BBB. The reset switch is linked to the Energy measurement field. As energy usage is power consumption integrated over time it needs a starting point in time. The Reset button gives a command to the BBB to restart the energy consumption calculation. The energy savings are also calculated by integrating the difference between the current power measurement and the previous. Further details will be discussed in section VI. Relay 1 and Relay 2 are used to send a command to the power measurement circuit to switch the respective relays on or off. Another relay switch could be added to the dashboard as the hardware made provision for three relay switches to control the devices' power supply

VI. SOFTWARE DESIGN

A simplified version of the main control loop for the system is shown in Fig. 5.

This control loop consists of an infinite while-loop that waits for certain events to take place. Within every cycle it checks if a command has been scheduled from the platform and if a positive result is found it determines what command has been sent and then executes that command. The rest of the events are linked to timing. Every 10 seconds the EMIC is asked for a power measurement which is integrated over time to calculate the energy consumption. This power measurement is also compared to the previous measurement and if it is smaller the potential energy savings are also calculated. These two values are then stored and can be cleared by the user by setting the reset switch to "on" which will reset the values once. After 30 seconds have passed the voltage and current readings are requested from the EMIC and are sent to the platform with the energy consumption and potential energy

saving values. The reason for the these two different timing intervals is due to the fact that power consumption in kWh needs to be calculated at short intervals by integrating the instantaneous power consumption over time, while it is not necessary to update the instantaneous voltage, current and power measurements so frequently

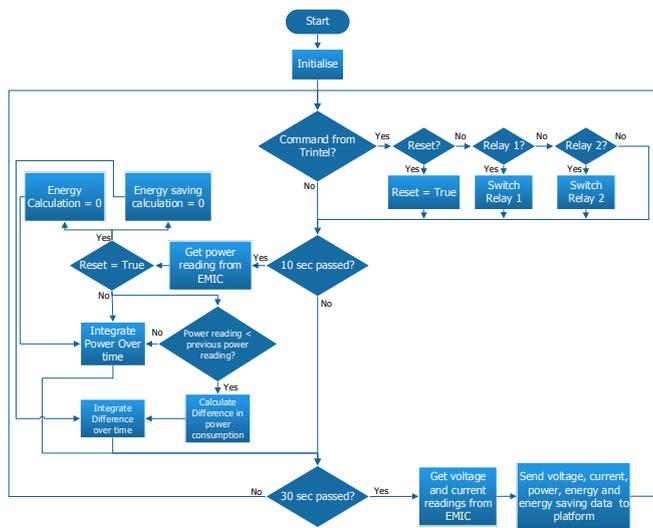


Fig. 5. Main Program Flow Diagram

A. Smart Platform Communication

The communication between the Beaglebone and the platform was performed using the BeagleBone’s UART1.

Table 1 lists the different type of commands used in this project.

Table 1: AT commands used in this project

Command	Description
at+cfun=1	Restarts the modem
at+boot?	Reboots the modem
at+cgsn	Requests modem’s serial nr
at+awtda=c*	Instructs modem to send all data back that it does not understand
at+awtda=d*	Instructs modem to send all commands back it does not understand
at+awtda=d, "Home_contr ol.Measure", 1, "V, INT32, 230"	This format is used to send the data variables to the SMART platform.
at+awtda=a, "Home_Contr ol", 1018977	This format sends job acknowledgements to the platform.

In order to receive commands from the platform a function `uart1_read()` was written. When this function is called it reads and processes the command from the platform to determine what type of command it is. Once the command type has been determined the command is executed and an acknowledgement message containing the commands unique ID number is sent back to the platform. The communication between the BeagleBone and the Platform consisted solely of data messages and the function

`uart1_write()` was used to send this data. The `uart1_write()` function receives the data that has to be sent and converts it to a string format that the platform can interpret using a `String_Creator()` function.

B. Power Sensor Communication

The EMIC responds to host commands that are sent from the BeagleBone’s UART2 port. There are four of these host commands needed to read and write data to and from the EMIC’s registers as well as to give instructions to the calculation engine. The different types of commands are listed in Table 2: **Binary Structure for communicating with EMIC** found at [21].

Table 2: Binary Structure for communicating with EMIC

Function	Binary Value	Description
Register Read	00 A5 A4 A3 A2 A1 A0	A[5:0] specifies the register address
Register Write	01 A5 A4 A3 A2 A1 A0	
Page Select	10 P5 P4 P3 P2 P1 P0	P[5:0] specifies the page number
Instruction	11 C5 C4 C3 C2 C1 C0	C[5:0] specifies the instruction

To retrieve the power measurements from the EMIC two commands are sent to it using the `uart2_write()` function, namely Page Select and Register Read. These two commands tell the EMIC on what page and register the requested data is stored in respectively.

Once the EMIC has received a command to return a specific measurement it will send three bytes of register data to the BBB. This data is sent using a total of 10 bits per byte with one start bit, eight data bits, and one stop bit with the least significant bit first. The values that the EMIC sends to the BBB are in the range of -1 to 1 and this value has to be converted to the actual value by multiplying it with a factor that is determined during calibration. The function `uart2_read()` is used to read and convert the data to a most significant bit first format.

In order to calculate the energy consumption of the monitored device, the power measurement is retrieved every 10 seconds, as seen in Figure 5, and then integrated over time to get the energy consumption, E , in kWh. It is calculated using Equation 1.

$$E = \sum (P \times 10s \times \frac{1kW}{1000W} \times \frac{1h}{3600s}) \quad (1)$$

This calculation is continually executed until the user triggers the reset command at the terminal which will restart the process. The calculation process of the potential energy savings is started by comparing the current power measurement with the previous power measurement. If the current power measurement is at least 10% smaller than the previous measurement a counter is incremented. This process is repeated three times resulting in a delay of 30 seconds to make sure that a load has been switched off and that it wasn’t

just a dip in the power consumption. When the counter reaches three the difference in power consumption is calculated by integrating over time using Equation 1. This integration process will keep on executing until the current power consumption becomes more than the previous power consumption or until the user triggers the reset command.

VII. CONCLUSIONS AND RECOMMENDATIONS

The completion of this project was dependent on finishing two parts, namely the power measurement and platform communication systems. The power measurement circuit performed well as it communicated successfully with the Beaglebone and was able to measure the voltage, current and power to within an accuracy of 0.6%. This accuracy is attributed to the CS5490 IC that was used for the power measurements and was definitely a good choice for a project of this nature.

The communication with the Smart Platform was successful as data and commands could be sent to and from the platform enabling easy access to the power consumption measurements. The Trinity Smart Platform was perfectly suited to this project, it worked very well as it allows for complete customization, enabling a user interface to be created which is specific to this project. It proved to be an effective M2M platform.

The Beaglebone Black was chosen as the MEC and proved to be a very capable device and a lot was learnt from it. The ability to develop capes which can fit onto the Beaglebone's headers makes it a very customisable device. This also enables faster design and development of circuits that can interact with the Beaglebone and makes the possibilities of implementation endless.

One of the biggest problems found was the speed at which the power measurements could be sampled in order for energy usage to be calculated. With the EMIC's UART set to the default 600baud the updating time was limited. It is recommended that the baud rate be set higher so that the energy consumption can be calculated at closer intervals which would increase the accuracy of the calculation.

The calibration of the EMIC also proved to be a bit of a challenge as its accuracy can only be as good as the device used to calibrate it, as calibrating it with a normal multimeter was not sufficient. It is recommended to use a proper energy measurement device as this will significantly increase the ICs' accuracy.

To improve the effectiveness with which this system can help users save electricity a future iteration of this project should include not only the potential energy savings data but also the quantification of money that is being saved, thus re-emphasizing the incentive of reducing energy consumption.

ACKNOWLEDGMENT

The authors would like to thank MTN and Trintel for financially supporting this project.

REFERENCES

- [1] Eskom, "49m", 18 03 2014 [Online]. Available: <http://www.49m.co.za/about>. Accessed 09 05 2014
- [2] Eskom, "Eskom-Integrated Demand Management," Eskom, 2010. [Online]. Available: <http://www.eskom.co.za/sites/idm/?Pages?Home.aspx>. [Accessed 09 05 2014].
- [3] Eskom, "AboutElectricity," [Online]. Available: http://www.eskom.co.za/AboutElectricity/FactsFigures/Documents/GX_0001GenPlantMixRev13.pdf. [Accessed 24 05 2014].
- [4] T.Lodge, "The RDP: Delivery and Performance," *Politics in South Africa: From Mandela to Mbeki*, 2003.
- [5] energy.gov, "petroleum," 1998. [Online]. Available: http://www.energy.gov.za/files/resources/petroleum/wp_energy_policy_1998.pdf. [Accessed 24 05 2014]
- [6] Beeld, "Noodstaple oor krag," *Beeld* 19 02 2006
- [7] News24, "The-People-in-power," 30 01 2008. [Online]. Available: <http://www.news24.com/xArchive/News24/The-People-in-Power-20080130>. [Accessed 24 05 2014].
- [8] IOL, "go-to-sleep-early-to-save-power," [Online]. [Available: <http://www.iol.co.za/news/politics/go-to-sleep-early-to-save-power-minister-1.387594#.U4B3QP15Xnc>. [Accessed 24 05 2014]
- [9] Eskom, "The Eskom National Efficient Lighting Program," [Online]. Available: http://www.eskom.co.za/OurCompany/SustainableDevelopment/ClimateChangeCOP17/Documents/The_Eskom_National_Efficient_Lighting_Programme_Compact_Fluorescent_Lamps_Clean_Development_Mechanism_Project.pdf. [Accessed 24 05 2014].
- [10] Natpower, "Geysersblanket," [Online]. Available: <http://www.natpower.co.za/pdf/geysersblanket.pdf>. [Accessed 24 05 2014].
- [11] AMEU, "library- Industry documents," [Online]. Available: <http://www.ameu.co.za/library/industry-documents/nrsubs/DSM%20ESLC%20information%20Rev%202.pdf>. [Accessed 24 05 2014].
- [12] Engineeringnews.co.za, "Eskom Energy Efficiency," 23 07 2010. [Online]. Available: <http://www.engineeringnews.co.za/article/eskom-energy-efficiency-2010-07-23>. [Accessed 24 05 2014].
- [13] Engineeringnews.co.za, "Eskom-places-temporary-hold-on-energy-efficiency-rebate-programmes," 09 12 2013. [Online]. Available: <http://www.engineeringnews.co.za/article/eskom-places-temporary-hold-on-energy-efficiency-rebate-programmes-2013-12-09>. [Accessed 24 05 2014].
- [14] Yahoo News, "south-africas-eskom-says-power-supply-severely-constrained," [Online]. Available: <http://news.yahoo.com/south-africas-eskom-says-power-supply-severely-constrained-141324787--finance.html>. [Accessed 24 05 2014].
- [15] Timeslive, "risk-of-blackouts-very-high-eskom," 10 01 2012. [Online]. Available: <http://www.timeslive.co.za/local/2012/01/10/risk-of-blackouts-very-high-eskom>. [Accessed 24 05 2014].

- [16] Energy.gov, “r_solar,” [Online]. Available: http://www.energy.gov.za/files/esources/renewables/r_solar.html. [Accessed 24 05 2014].
- [17] Evwind.es, “solar-energy-in-south-africa,” 11 04 2012. [Online]. Available: <http://www.evwind.es/2012/04/11/solar-energy-in-south-africa/17693>. [Accessed 24 05 2014].
- [18] Africa Energy Intelligence, “Pan African Consortium for Inga,” 2003.
- [19] bdlive.co.za, “das-economic-plan-targets-8-growth,” 18 02 2014. [Online]. Available: <http://www.bdlive.co.za/national/politics/2014/02/18/das-economic-plan-targets-8-growth>. [Accessed 24 05 2014].
- [20] Sparkfun, “RS-232 vs TTL serial communication,” [Online]. Available: <https://www.sparkfun.com/tutorials/215>. [Accessed 04 05 2014].
- [21] Cirrus Logic, “Two Channel Energy Measurement IC,” 2013.
- [22] S. Wireless, “Sierra Wireless Airvantage Platform,” 2010.
- [23] T. Telecomms, “www.trintel.co.za,” Trintel, [Online]. Available: <http://www.trintel.co.za/content/4020/0/home>. [Accessed 01 05 2014].

VIII. BIOGRAPHIES

Douw du Plessis received the BEng degree in Electrical and Electronic Engineering from Stellenbosch University in 2014.

Presentation of a Home Automation Solution with Potential for Seamless Integration and Vast Expansion

G. Sawyer

Department of Electrical and
Electronic Engineering
Stellenbosch University
Western Cape, South Africa
Email: 15655903@sun.ac.za

M.J. Booysen

Department of Electrical and
Electronic Engineering
Stellenbosch University
Western Cape, South Africa
Email: mjbooyesen@sun.ac.za

Abstract—The ever-increasing existence of electronic systems and devices within the residential environment, along with the human desire to simplify life and daily routine, is generating increased interest in the field of Home Automation and intelligent environments. A large variety of HA solutions have been conceptualised or developed. However, many of these solutions are designed by experts and therefore require professionals to install and/or operate them. Furthermore they lack the potential for seamless integration into an already functioning home environment. This paper presents a HA solution with seamless integration potential. The system can be installed and configured without professional skills or physical alteration of the environment itself. There is also large potential for the expansion of the systems capabilities and functions due to the hardware and software platforms utilised. This paper concludes with an analysis of performance tests results, and a discussion of the potential avenues for expansion.

I. INTRODUCTION

Home Automation (HA) is the integration of electrical and electronic devices within the residential household, via a centralised control system, which provides the habitant with control over the devices, either locally or remotely, as well as the ability to automate certain processes or functions within the home [1]–[3]. For a system to be considered a successful HA system, it must integrate a number of application areas. These application areas are listed below:

- Device function automation
- Intelligent control
- Power consumption management
- Sensor rich environment
- Sensible user interface

HA has received a lot of attention recently due to the ever increasing number of electronic and electrical devices within the home. Along with an increased interest in Machine-to-Machine communication networks, there is a lot of support in the field of HA. The ability to host communications between a number of devices and a centralised intelligent control system, opens the door to a lot of potential for growth [4]. An area of HA which has received significant interest recently, has

been that of augmenting HA networks with a multitude of sensors. An environment which hosts the integration of devices and sensors in an interconnected network with an intelligent control system, is known as an intelligent environment. With the addition of communications and an awareness of a humans presence and location, users could be empowered by a digital environment that is sensitive, responsive and adaptive to their behaviour and needs [5], [6].

A. Home Automation Currently

1) *Driving Factors*: One of the largest driving factors behind HA is the human desire to simplify and improve life in the home environment. Intelligent environments create a lot of potential for cutting the cost of living by removing the need for human intervention in power saving routines that would otherwise seldom be performed (such as turning the hot water cylinder off when leaving the home). Safety and security would also benefit by, for example, the use of CO₂, gas and motion sensors coupled with alarm and notification systems. These are but a few areas in which HA can offer benefits to not only the occupant, but also the industry upon which the occupant relies.

2) *Challenges*: One of the largest challenges to HA is overcoming the lack of consumer support. In the eyes of the consumer, automating the household is an expensive and unnecessary endeavour. The HA systems available today are purpose-built to spec for each home and therefore require professional intervention to install and setup. This inflates the cost of HA and also creates the illusion that it is intended as a luxury, ie. used with media and entertainment alone. This illusion along with the high cost of these systems drives consumers away and this has resulted in a lack of positive end-user generated data (or consumer sentiment) to support the financial advantages, and other benefits of HA. The solution proposed in this paper addresses this issue directly as it is not only cost effective, but it integrates each of the application areas that constitute a HA system. Also, it is capable of

seamless integration into any home without the need for professional assistance resulting in a lower cost, and the ability to easily reconfigure the system as the home environment evolves.

On the other hand, design of an HA system with seamless integration potential is challenged by the lack of a universal development platform. Currently, most HA solutions are developed as a closed box. For this field to progress in the future, focus must be placed on integration with a large variety of systems and devices. The solution proposed in this paper addresses this challenge as it serves to be a HA platform upon which there is much room for expansion.

3) *Generic HA Network:* The generic architecture for a HA network consists of sensors/actuators, an intelligent controller and a gateway through which the network can access the internet or be accessed by a user. The sensors and actuators are deployed throughout the home to monitor and control the in-home environment. Information is communicated to and from the sensors/actuators by the intelligent controller. The controller provides sensor information to the user and also allows the user to carry out control actions on devices within the home. The intelligent controller is capable of independently carrying out control actions based on events monitored and translated into triggers. Fig. 1 depicts how the ideal HA architecture might look when connected to a variety of devices, using a variety of network technologies.

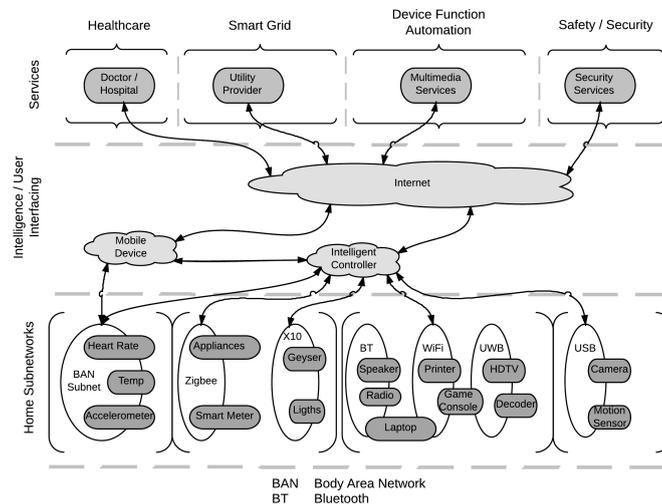


Fig. 1. Generic Home Automation Architecture

B. Contribution

This paper presents the design and development of an easy to use, highly expandable and seamlessly integratable, proof-of-concept Home Automation solution. The system provides users with the ability to automate appliances and processes within their home without the need for complex instalments and connections. The goal can be to lower power consumption of a hot water cylinder or to improve safety, it is up to

the user how the system is deployed. Interfacing the user with the system is designed to be effortless. A web server, hosted locally on the system, provides device and sensor status information to the user allowing them to monitor the status of their home and carry out control of the connected devices. The web-server can be accessed from a remote location over the internet, and status information is automatically pushed to the user and updated live. In the event a user decides to change the device connected to a particular switch, they are also able to change the name of the device on the user-interface and this information is retained, even after a power outage. Lastly, on top of the ability to control devices directly, it is possible to generate intelligence based control loops. An event (such as temperature or time) can be used to trigger a control output (such as a light switch or geyser control). Furthermore, a control output can also be triggered using two events that are linked by a logical AND or logical OR.

The proof-of-concept presented in this paper is small in terms of its array of sensing and control abilities. But it has been developed to serve as a platform upon which expansion is made simple. With a modular design, the addition of more complex control functions and sensor capabilities requires little work.

C. Document Layout

The rest of this paper is organised as follows. Section II presents a brief survey of the literature on HA. Some of the important application areas are discussed along with proposed HA solutions. Challenges facing the progress of HA are also presented. Section III proposes a HA solution addressing the drawbacks of solutions proposed in literature. Section IV presents the design of the HA solution proposed and finally section V concludes this paper.

II. LITERATURE SURVEY

HA has received substantial attention in the past decade or so and as a result there is a lot of research to support all areas of the field. The application areas that have received the most attention are discussed as well as some of the solutions proposed by researchers in the field. This survey gives us a clear indication of some of the important challenges that face the future of HA.

A. Most Common Home Automation Application Areas

The typical HA network is comprised of appliance devices, sensors, smart grid components, healthcare devices and security systems. The goal of a HA network is to improve the quality of life, experienced by the user, by performing tasks based on intelligence which holds to a set of rules. These tasks can be anything from toggling lights when motion is detected, to preparing the home for vacancy when the users go on holiday. In order to realise a HA system that can fulfil this goal, focus must be placed on key application areas. Some of these application areas are discussed here.

An important application area is that of device automation (such as lights, heaters, appliances etc.). The benefits that come

with device automation are obvious, and consumers notice these benefits easier than those in other application areas. This area concentrates on devices carrying out functions without intervention from the user, and allowing the user to control these devices, over the Internet, from outside their home [7]. Much of the research in the field of HA is dedicated to device automation [4], [8], [9].

Smart grid implementation is one of the predominant focuses of HA. This is because smart grids will not only assist the consumer financially, but will also allow for much better provision of services by their providers [10]. The main objectives of smart grids are to increase the efficiency of power transmission, increase the quality of service to utility users and to reduce the economic and environmental cost of power generation and consumption. With a HA system a user is able to monitor their homes energy usage as well as limit the consumption of particular devices. Providing this information to the utility generators would ultimately give them the ability to provide the consumer with energy on demand, as well as offer them benefits based on consumption.

Intelligent environments is another of the most common application areas. An intelligent environment is defined as one where computing technology is embedded within the environment in such a way that it becomes virtually invisible [11]. Through the use of M2M communications and intelligent control systems, as well as an awareness of a humans presence and location, the aim is to empower users with a digital environment that is sensitive, adaptive and responsive to their behaviour and needs. Most modern day automation systems are designed to carry out a predetermined process [1], [12], triggered by an event; such as time, user action, a sensor reading or the result of a previous process (by process we mean actuation of a device in the home such as switching lights or changing the temperature of the air conditioner). The aim of developing an intelligent environment is to eliminate user interference as much as possible and increase the benefits to the user living within the environment [13].

B. Research to Support Home Automation Development

HA consists of a wide array of application areas and while a lot of effort is put into researching these areas individually, there is not a lot of focus on a HA system that incorporates everything. A capable HA system must integrate with all types of hardware within the home, and allow for addition and removal of devices to and from the network. That is to include devices with logic and communications, such as a smart television, and also devices without, like a toaster or lights. It must be designed with a user friendly interface that is easily accessible while also aiming to lower user intervention as much as possible.

The most common application of HA is that of the appliances and devices within the home. The switching of lights based on occupancy or preparing the kitchen for the morning routine are both attractive capabilities of HA however, no two homes use the same appliances, nor do they use them in the same manner. The challenge is therefore to develop a

HA system capable of integrating seamlessly into any home no matter the appliance type or function. In [12] and [7] similar systems capable of interfacing with electronic hardware in a HA system are proposed. In [12] an Arduino development micro-controller is used to drive relays that are connected to electronic devices, and control is carried out via a smartphone with Bluetooth connection. In [7], some slightly more complicated, purpose-built hardware was used but the end result was much the same.

In [10] the authors present a proof-of-concept system to monitor and control household water heating on a large scale, using a web based interface. The system gives a user the ability to monitor and alter the HWC temperature and also safely empty the HWC in the case of failure or maintenance. The pressure of the HWC is also monitored to detect failures.

C. Challenges Facing Future Progress of Home Automation

HA is faced with many challenges hindering future progress, one of which is the standardisation of the technology being used. The home environment is populated by a wide variety of devices and systems all being controlled by a multitude of communication protocols, and producing a vast array of information types. Many concepts have focused on developing software capable of handling the broad variety of data when dealing with HA networks [9], [14] while others have focussed on hardware capable of integrating with many different devices. The main challenge to HA is incorporating all the application areas into one solution. The research mentioned in the previous section along with other existing solutions focus on one aspect of HA but fail to consider how they might all be integrated with each other. From here we go on to propose a HA proof-of-concept, providing the functions detailed in the contribution, capable of addressing these challenges.

III. DESIGN REQUIREMENTS

An expandable and seamlessly integratable HA proof-of-concept is proposed. The system must be capable of interfacing with simple electrical appliances as well as devices that require some communication protocol (such as IR). Expansion and community development must be possible to prevent limitation to a small group of unique users. The system must also make use of a modular design allowing users to add and remove devices to and from the system since no two home environments are the same. Lastly the system must give a user the ability to control devices directly, and also to generate intelligent rules that govern the control of devices based on sensor or event data.

A. Functional Requirements

Key factors that are to be addressed in the design of the proof-of-concept are; cost effective and open source hardware/software, slave hardware with modular capabilities, master software with adaptive capabilities, non-intrusive and off-the-shelf installation, and an easy to use and universal user-interface.

B. System Requirements

Hardware used in this HA system must be cost effective to keep the end user cost low. It must also run open source software to allow for vast expansion and support potential from the developer community. For a fully integrated HA solution to succeed, it cannot only rely on a small group of paid support staff. As mentioned earlier, each user may want their HA system to behave in a different way to the next. This requires the support of the open source community.

This concept proposes a master, slave configuration for the HA proof-of-concept. This means that slave nodes are placed throughout the home, and each one reports information to, and receives commands from a master node. The master node handles communication to and from the slaves, as well as interfacing with the user. This allows the incorporation of a modular slave design, and gives the ability to add and remove slaves from the system seamlessly. Furthermore, slaves can be customised to perform a unique set of functions or report a unique set of sensor readings, and added to the system without the need to alter the system itself. The software on the master node must be adaptive. It must be able to detect when new slaves are added, and when they are removed and it must also be able to identify these slaves and the devices they are connected to.

The HA solution must not require that wiring and other electrical installations need to be embedded within the walls of the home. It must also not leave wires running throughout the home in a fashion that it becomes a burden on the user. For this reason, all communication between slaves and master must be wireless. Slaves must also require as little wiring as possible (such as between the slave and an array of relay switches).

The user interface must be simple to use with distinct sensor outputs and control inputs. It must also be accessible to all users whether they are using a mobile phone, tablet, or pc (running any operating system).

IV. SYSTEM DESIGN

This section details the design of the HA proof-of-concept solution based on the requirements set out in the solution proposal.

A. System Layout

Fig. 2 shows the system layout. Blue arrows indicate information flow between slave nodes and master node. This information consists of sensor readings and control commands. Red arrows indicate flow of information between the user and the master node. This information consists of the providing the user with the web server interface, as well as allowing the user to view sensor reading and perform control commands. Using this system layout, the user can access their HA system either by connecting directly to their home gateway, or by accessing it through the internet from anywhere in the world. Also, if internet connectivity is dropped, the system will continue to function and can still be accessed with a local connection.

This system layout also takes full advantage of the proposed modular slave design. If one slave goes down, the system will

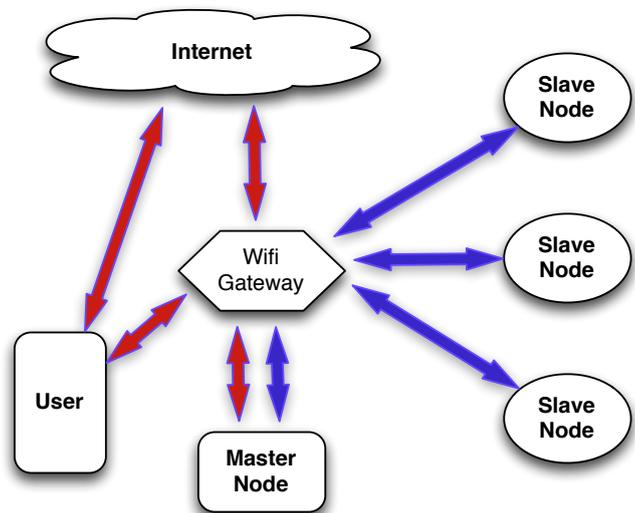


Fig. 2. System Layout

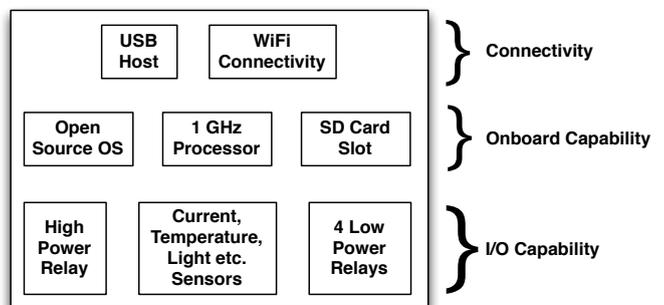


Fig. 3. Functional Slave Layout

continue to function without a disruption of communication. Also, slave can be added and removed without the need to reset the system. More detail will be given on how this is achieved in the section on software.

B. Hardware

Hardware choice is of large importance in the success of this proof-of-concept and consists of the selection of a micro-controller to serve as the master and slave nodes. The hardware must be cost effective, but also be capable of fast, wireless communication (using IP to facilitate the modular system design where slaves are treated as clients) and processing of accurate sensor information that may require a fast CPU. Also, the selected component must be open to third party development in order to promote the growth and expansion of this concept as well as HA in general. Hardware options were considered after deciding on the required capabilities of the slave node. Fig. 3 shows a layout diagram of the functional capability of the slave node.

Based on these functional requirements, the Beaglebone Black [15], made by CircuitCo [16] was selected. The Beaglebone Black has the processing power required, as well

as a multitude of input/output pins capable of digital and analog I/O and many forms of serial communication. The Beaglebone Black is also compatible with Arduino shields making it open to a wide variety of expansion. Lastly, the Beaglebone Black is capable of running Linux, Android and Ubuntu making it a suitable development platform.

C. Software

Three main design areas were taken into consideration when choosing the platform on which to run the software, along with the language in which to program the software. The first, and main consideration is that of the modular design requirement. After extensive research into the available software packages and languages, as well as ongoing projects in this field, it was discovered that an ongoing community project existed with software capable of fulfilling this requirement. The developers of The Thing System [17] designed a software package called the Steward [18]. The Steward connects the things (devices such as lights and electrical appliances) in the home, whether those things are media players such as the Roku or the Apple TV, a Nest thermostat, INSTEON home control system, or the Philips Hue lightbulbs whether these things are connected together via Wi-Fi, Zigbee, Z-Wave, USB or Bluetooth LE. The Steward is capable of identifying these things, and allowing them to talk to one another. What's more important about this software package is that devices can be programmed to communicate with the Steward API using any language. Also, the Steward API is open source and extensively documented. However, the Steward cannot currently interface with devices that have not been designed to communicate with the Steward API or are not equipped with any communication protocol at all. This proof-of-concept design makes use of the Steward for its capabilities as well as add the ability to integrate the simple devices, that exist in within the home environment, to the automated network. Fig. 4 depicts the master/user and master/slave flows of information, while Table I details the specific functions carried out by the browser, master software and slave software.

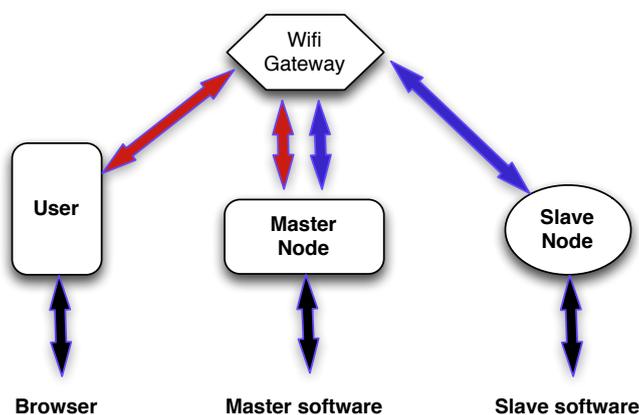


Fig. 4. Functional Software Layout (note. Only one slave is depicted in this figure).

TABLE I
DESCRIPTION OF FUNCTIONAL SOFTWARE COMPONENTS.

Browser	Master Software	Slave Software
- Requests web interface from client.	- Runs Steward software	- Aggregates sensor readings and reports to Steward.
- Allows user to view status informations and carry out control operations	- Runs client software that provides web interface to user.	- Carries out control functions received from Steward.
- Allows user to create, monitor and delete intelligent rule.	- Receives sensor statuses from Steward.	
	- Receives control commands from user.	

The second consideration was the language in which to write the master and slave nodes. The Steward is written in Node.js [19], a portable and extensible open source language. To keep things uniform, both the master and slave nodes will be written in Node.js, and at the same time, this is desired as it will help to promote the growth and development of this concept, and HA in general, within the open source community. The master and slave nodes must also be capable of bi-directional communication. Node.js proves to be a good language choice as there is a large support community constantly developing packages for use in Node.js. One of these packages supports WebSocket [20] communication in Node.js. WebSocket is a protocol providing full-duplex communications channels over TCP connection. This protocol allows a communication channel to be opened, and remain open for the duration of two-way communication.

The third consideration was the operating system on which to run the master and slave software. In keeping with an open source and expandable design, Debian [21] is selected as a suitable operating system. It has a large support community as well as the ability to run on development micro-controllers.

D. Slaves: Sensors and Actuators

Slave nodes are required to capture sensor readings and perform actuator control in specific locations within the home. The slave nodes have been designed in such a way that one slave node is required per living area (in a reasonably sized residential home) and have been equipped with some standard sensor and control electronics. Each slave nodes is capable of monitoring motion, temperature (to within half a degree C), light and current. Sensor reading are taken once every second and subsequently reported to the master node. Each slave is also capable of switching one high power relay for use with any device requiring up to 10A (in the event more current is required, this relay can be used to switch a higher power relay), three standard relays capable of up to 5A and one dimmer switch capable of a maximum of 300W. Furthermore, the standard relays and dimmer switch are wireless radio frequency controlled. This way we are keeping with the non-intrusive requirement.

E. Master: User Interaction

The master node simultaneously runs the Steward and client software. The client software is responsible for communicating with the Steward to provide sensor information, and receive control commands. The client is also responsible for running a web server which is provided to the user when they wish to monitor or control their home environment. The user-interface is designed to be easy to use, and universal across all platforms ie. mobile, tablet and desktop computer, and is accessed by pointing a browser to the IP address of the master node, and port of the web server. The user-interface also makes it easy to understand how the elements on the web page are linked to the physical devices within the home. This is done by allowing the user to customise device names upon connecting them to the system, or when interchanging the connected devices. Device name info is stored in an SQL database such that it is retained, even in the event of a power outage.

F. Intelligence

A crucial area of HA is intelligence. Intelligence means that a HA system must be capable of carrying out control functions based on sensor readings without intervention by the user. Along with the standard monitoring and control provided by the user interface, there is also capability for the creation of rules. In this area a user selects what sensor event they would like to monitor and what control function should be performed should that event occur. Sensor events include temperature or motion readings, as well as time. Events can also be linked together with a logical AND or a logical OR to create a more complex output control scenario. For example one may wish to switch the lights on in a particular room when motion is detected in that room and the ambient light level is below a certain threshold.

V. CONCLUSION

A. Proof-of-Concept Comparison

The resulting Proof-of-Concept realises a flexible HA system capable of integrating with all household devices and successfully achieving all the stated goals. This system differs, in a few key areas, from similar systems that were mentioned earlier. In [7] a similar HA solution is proposed. That system also makes use of a Master/Slave hardware setup, and also uses a web server to interface a user with the system. The largest difference is the use of a computer to store the database information and run the web server. The system proposed in this paper stores database information and also runs the web server on the master node. This results in both a lower complexity and cost of this Proof-of-Concept. The system proposed in [7] also makes use of RF communication between master and slave nodes, whereas the solution proposed in this paper utilises IP WiFi communication. WiFi, instead of RF, was implemented to allow for seamless expansion of system capabilities. The use of IP allows third party developers to produce and integrate slave nodes with a variety of capabilities into the HA system without needing to make changes to the master node of the system. The use of RF would require

alterations to the master node each time a new slave was introduced to the system.

In [12] a similar but slightly less advanced system is proposed where an Arduino Bluetooth board is connected, via digital I/O pins, to devices within the home and controlled by a Bluetooth capable cellphone running a custom python script. There are a few large differences between this system and the system proposed in this paper. The system in [12] would require changes to the script each time a new device is added to the system. It is also not capable of providing a user with remote access due to the use of Bluetooth communication. Finally, it is not capable of automation but instead serves as an appliance remote control.

B. Future Work

The proposed HA solution holds a lot of potential for future expansion. It is recommended that the work be carried forward by developing the integration of more complex devices such as those requiring infrared control. Along with this added integration, it will be necessary to develop the user interface further. The user interface will need to be capable of adapting to whatever variety of devices are connected. When devices are added or removed, it should automatically reflect on the UI and, sensor information and control capabilities should also automatically be presented to the user without needing to conform to a predetermined layout. Lastly, an investigation into the use of CoAP communication protocol should be carried out in order to increase energy efficiency of the system.

C. Conclusion

This paper presents the design of a solution to HA integrating all the important application areas. Focus is placed on developing a system capable of vast expansion and reconfigurability addressing some of main challenges of HA. Also, the system is designed to be simple to install and operate to promote consumer support of the technology. The paper discusses existing HA research and solutions identifying the drawbacks to these solutions and the challenges facing the field. A concept is proposed that addresses the challenges and is capable of compensating for the drawbacks of existing solutions. A detailed design of the system is presented and the capabilities and potential of the solution is discussed.

ACKNOWLEDGMENT

G Sawyer and MJ Booyesen were funded by MTN.

REFERENCES

- [1] S. Folea, D. Bordenca, C. Hotea, and H. Valean, "Smart home automation system using Wi-Fi low power devices," in *Proceedings of 2012 IEEE International Conference on Automation, Quality and Testing, Robotics*. IEEE, May 2012, pp. 569–574. [Online]. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6237775>
- [2] F. Hamoui, C. Urtado, S. Vauttier, and M. Huchard, "SAASHA: A Self-Adaptable Agent System for Home Automation," in *2010 36th EUROMICRO Conference on Software Engineering and Advanced Applications*. IEEE, Sep. 2010, pp. 227–230. [Online]. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5598101>

- [3] N. Liang, L. Fu, and C. Wu, "An integrated, flexible, and Internet-based control architecture for home automation system in the Internet era," in *Proceedings 2002 IEEE International Conference on Robotics and Automation (Cat. No.02CH37292)*, vol. 2. IEEE, 2002, pp. 1101–1106. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1014690
<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1014690>
- [4] C. Felix and I. Jacob Raglend, "Home automation using GSM," in *2011 International Conference on Signal Processing, Communication, Computing and Networking Technologies*. IEEE, Jul. 2011, pp. 15–19. [Online]. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6024506>
- [5] U. Bischoff, V. Sundramoorthy, and G. Kortuem, "Programming the smart home," in *3rd IET International Conference on Intelligent Environments IE 07 (2007)*, 2007, pp. 544–551. [Online]. Available: http://digital-library.theiet.org/content/conferences/10.1049/cp_20070424
- [6] K. Wang, W. Abdulla, and Z. Salcic, "Multi-agent fuzzy inference control system for intelligent environments using JADE," no. 2, pp. 285–294, 2006. [Online]. Available: http://digital-library.theiet.org/content/conferences/10.1049/cp_20060653
- [7] A. Alkar and U. Buhur, "An internet based wireless home automation system for multifunctional devices," *IEEE Transactions on Consumer Electronics*, vol. 51, no. 4, pp. 1169–1174, Nov. 2005. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1561840
<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1561840>
- [8] M. Bjelica and N. Teslic, "A concept and implementation of the Embeddable Home Controller," *MIPRO, 2010 Proceedings of the 33rd ...*, pp. 686–690, 2010. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5533494
- [9] M. Z. Bjelica, I. Papp, N. Teslic, and J.-M. Coulon, "Set-top box-based home controller," in *IEEE International Symposium on Consumer Electronics (ISCE 2010)*. IEEE, Jun. 2010, pp. 1–6. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5523704
<http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5523704>
- [10] M. J. Booysen, J. A. A. Engelbrecht, and A. Molinaro, "PROOF OF CONCEPT : LARGE-SCALE MONITOR AND CONTROL OF HOUSEHOLD WATER HEATING IN NEAR REAL-TIME," 2013, pp. 1–8.
- [11] F. Doctor, H. Hagrais, and V. Callaghan, "A Fuzzy Embedded Agent-Based Approach for Realizing Ambient Intelligence in Intelligent Inhabited Environments," *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, vol. 35, no. 1, pp. 55–65, Jan. 2005. [Online]. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1369345>
- [12] R. Piyare and M. Tazil, "Bluetooth based home automation system using cell phone," in *2011 IEEE 15th International Symposium on Consumer Electronics (ISCE)*. IEEE, Jun. 2011, pp. 192–195. [Online]. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5973811>
- [13] L. Zhang, H. Leung, and K. Chan, "Information fusion based smart home control system and its application," *IEEE Transactions on Consumer Electronics*, vol. 54, no. 3, pp. 1157–1165, Aug. 2008. [Online]. Available: <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4637601>
- [14] A. Alvi and D. Greaves, "A logical approach to home automation," *Intelligent Environments, 2006. IE 06. 2nd ...*, pp. 45–50, 2006. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4199363
- [15] (2014, 08). [Online]. Available: <http://beagleboard.org/>
- [16] (2014, 08). [Online]. Available: <http://circuitco.com/>
- [17] (2014, 08). [Online]. Available: <http://thethingsystem.com/>
- [18] (2014, 08). [Online]. Available: <https://github.com/TheThingSystem/steward>
- [19] (2014, 08). [Online]. Available: <http://nodejs.org/>
- [20] (2014, 08). [Online]. Available: <https://www.websocket.org/>
- [21] (2014, 08). [Online]. Available: <https://www.debian.org/>

An OpenMTC platform-based interconnected European – South African M2M Testbed for Smart City Services

Andreea Ancuta Corici¹, Asma Elmangoush², Thomas Magedanz¹, Ronald Steinke²
Fraunhofer FOKUS¹

Technische Universität Berlin²
Berlin, Germany

ronald.steinke@tu-berlin.de, asma.e.almangosh@campus.tu-berlin.de, thomas.magedanz@fokus.fraunhofer.de,
andreea.ancuta.corici@fokus.fraunhofer.de

Joyce Mwangama, Neco Ventura
University of Cape Town
Cape Town, South Africa

joycebm@crg.ee.uct.ac.za, neco@crg.ee.uct.ac.za

Abstract— Recent advances in device, information and communication technologies have exhibited a strong potential to enhance the quality of life of inhabitants of Smart Cities. Interconnecting different Smart Cities' infrastructures can help in the exchange of information and experiences between developed and developing nations. However, more research work is needed to define the aspect of interworking smart services, and evaluate the benefits archiving this. In the developed and the developing world, there are common use cases where existing solutions can be reused or easily adapted. Nevertheless, there are also different use cases which may require different test environments simultaneously for efficient testing. Additionally, different environmental conditions can be used for experiment aspects. In this paper, we introduce an interconnected M2M platform testbed in the Smart City context based on the standardized M2M platform OpenMTC. Our approach aims to interconnect testbeds in Germany and South Africa, and enable resource sharing between systems. This approach is framed in the context of the TRESIMO project. The integrated resources include sensors or devices used in the different test environments as well as applications. We present the planned experiments for the common and different use cases, which use devices located in either or both environments, with the aim of investigating possible problems arising due to delay and reliability. The experiments will consider different environmental settings to compare the considered scenarios in both developed and developing world.

Keywords— M2M; IoT; Smart City; Testbed

I. INTRODUCTION

In a Smart City, various application domains need to work together as an integrated large-scale infrastructure to cover multiple operations that this complex system needs to perform. These integration points include the communication network, the Internet, sensors, devices, gateways, and the resources or services of the Smart City. Different technologies will be involved to enable the Smart City implementation, and the

challenges towards the realization of smart services are numerous. Machine-to-Machine (M2M) communications, Wireless Sensor Networks (WSN) and other related technologies are the subject of many ongoing discussions in numerous academic disciplines and practical areas. Most studies highlight a limited set of factors from a variety of common components underlying a smart city [1][2]. One of the main challenges faced by different research directions is the lack of large-scale testbeds [3]. Most studies present a proof-of-concept using intended testbeds or systems, which might not be suitable for validating other concepts [4].

The heterogeneity of integrated devices adds another challenge when testing new services within the Internet of Things (IoT) support framework. More than nine billion devices around the world are currently connected to the Internet, and estimations show that by the end of 2020 there will be one trillion connected devices worldwide [5]. For service testing, the adjustment of many variables have to be taken into account which depends on the integrated devices or technologies.

The main contribution of this paper focuses on the design and implementation of a large-scale testbed for Smart City research. The testbed is based on a standardized M2M platform and open-source framework for managing and federating testbeds. Our experimentation framework will interconnect two M2M testbeds located in Berlin, Germany and Cape Town, South Africa, allowing the testing of multiple scenarios of Smart City services in both developed and developing countries. Additionally, the federation of both testbeds will allow the sharing of resources (i.e. sensors, actuators and data) between different services and users regardless of their location. The aim of the testbed is to support the investigation of M2M enabling technologies, architecture, devices and applications in

different environmental conditions. This paper presents the components of the proposed testbed and our approach in experimenting Smart City services.

The rest of this paper is organized as follows: in section II, the related work in the context of this paper is presented. Section III overviews the some scenarios of Smart City services that will be considered in the planned experimentation. Section IV includes a description of the designed large-scale testbed and the used components. Finally, the conclusion and future work is provided in section V.

II. RELATED WORK

A. Emergence of Smart Cities

Smart Cities are widely considered as a hot topic in academia and industry; however, there is no clear definition of the Smart City concept among practitioners. Authors in [6] reviewed several working definitions and proposed a general Smart City framework based on eight factors: “management and organization, technology, governance, policy context, people and communities, economy, built infrastructure, and natural environment”. The instrumentation of Smart Cities is considered as a key enabler, that will leverage the understanding of the city operations by “making the invisible visible” [7]. The following are key requirements that all Smart Cities must aim to achieve:

- Utilization of existing underlying communication infrastructure;
- Cost effective data storage and management;
- Systematic computational analysis of data or statistics;
- Interoperability of connected systems and services.

A number of research projects focused on Smart Cities deployment and experiments, such as SmartSantander [8], LOG-A-TEC [9] and I3ASensorBed [10] to provide valuable insights of testbeds on a controlled environment. However, their work lack the interoperability with heterogeneous systems and platforms.

B. TRESCIMO Architecture

The Smart City project, entitled “Testbeds for Reliable Smart City Machine-to-Machine Communication” [11], aims to address Smart and Green Cities challenges within underdeveloped countries. In this section we describe the reference architecture. The overall architecture is presented in Fig. 1, which was defined to fulfil the following objectives [12]:

- Deliver a specification of the overall architecture that involves an M2M communication platform used as the basis for a Smart City platform.
- Interweave a standard-based M2M platform with other sophisticated Smart City platforms.
- Integrate resource-constrained devices over Delay Tolerant Networks (DTN).
- Perform the integration of the main building blocks (M2M, Smart City, and Smart Energy) into a comprehensive platform using federation tools.

- Define specific enhancements for a Smart/Green City system, by implementing one trial for Smart Energy consumption in the Gauteng region (South Africa) and one trial for environmental monitoring in San Vicenç dels Horts (Spain).

The TRESCIMO architecture is realized by three-tiered layers, consisting of a sensors/things tier, an M2M gateways tier, and a service control tier. The sensor/thing tier provides the necessary mechanisms to digitize the physical data from the surrounding environment. Different types of devices and sensors can be connected to the system via suitable gateway interworking proxies. For power-constrained devices, a wakeup mechanism is developed as an energy-efficient solution for gathering data without real-time constraints [13].

The Gateway tier links the sensor/thing layer to the upper service control layer over standard open interfaces. The integrated M2M gateway is ETSI/oneM2M compliant, and supports multiple transport protocols like HTTP (IETF RFC 2616) and CoAP (IETF RFC 7252). The service control layer is based on a Smart City enabler platform, which has a REST-based data model to describe resources that expose data from connected entities and allows for both device discovery and actuation. The device management mechanisms are developed according to the oneM2M standard [14].

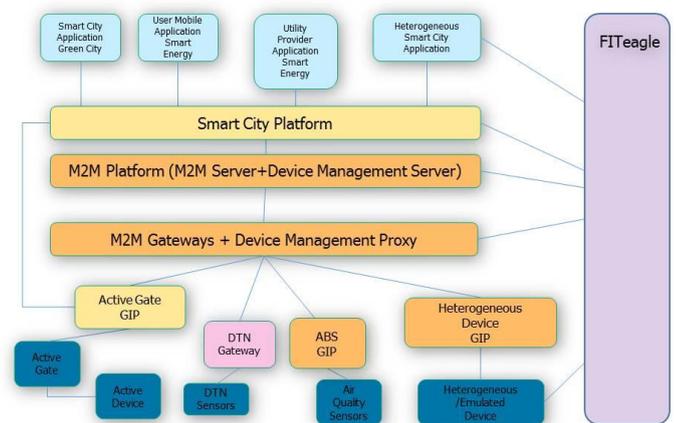


Fig. 1. TRESCIMO Reference Architecture

III. SCENARIOS AND EXPERIMENTATIONS

For the purpose of testing and experimentations, we consider some use cases that are commonly found in literature as key aspects of a Smart City. In this section, the addressed smart services and scenarios are described. Mainly, the goal of the experimentations is to perform the following tests:

1. Functionality testing: to validate the specified components of the system.
2. Configuration testing: to validate the performance of the implemented service under various environments, by conducting different measurements with configurable conditions.
3. Scalability testing: to demonstrate the ability of utilizing different hardware architecture or communication technologies in the designed solution.

A. Energy

The increasing demand of energy, depleting energy sources and the global warming problem caused due to the impact of energy usage are main issues that drive urban managers and operators to invest in studying technologies around the Smart City concept. Cities are looking to solve their problems with the development of new technologies to collect information and control energy in order to minimize urban energy consumption levels.

We define two scenarios for the Smart Energy use case; both involve an operator, customers and a utilities provider. In the first scenario, the customer is using a utility app that enables him to receive a daily report of his energy consumption. The analysis of aggregated information from the individual's energy consumption can help the operator in better understanding the customers' demands and usage pattern, and based on that offer suitable charging models. The customer shall benefit from such reports in learning about his/her usage pattern and for reduced consumption. The second scenario allows the operator to remotely control some devices that were assigned by their owners as "controllable". In case of energy supply drop due to some power station damage or maintenance, the power grid may become unstable and the operator would need to take action accordingly. First action shall be sending notification messages to customers located in the affected area to reduce their energy consumption. In the second step, the operator can send shutdown commands to some devices that were assigned by their owners as "controllable".

We are interested in investigating the performance of these scenarios using different communication technologies, and to experiment the efficiency of the control approach using lightweight device management protocols (such as OMA LWM2M) with the aim of maintaining the grid stability.

B. eHealth

Utilizing communication technologies in healthcare management aims to empower health specialists and improve the productivity of the service provided. However, in our test scenarios we will only consider informational services. The target is to help people be informed about any environment's status that may affect their health wherever they travel. For example, a person who has asthma will be concerned with the pollution levels in the air of the urban environment (air quality). Such a person would want to avoid the risk of an asthma attack. A simple application that retrieves this data from selected types of sensors (temperature, humidity, pollution, etc.) spread over the city, can be useful in order to highlight the areas of the city the individual should avoid when travelling around the city.

The data collected, by sensors measuring these environmental attributes, is usually small in size (few bytes) and generated at predictable frequencies. However, these sensors can be located in areas with poor communication coverage; there might be considerable delays in retrieving the information by the end-user. Our experimentations will consider various test environments to estimate the reliability of implementing similar (or more critical) eHealth services.

C. Transport

Intelligent Transportation Systems (ITS) are needed to support moving individuals (and goods) in an optimum time, and in a safe and cost effective manner. Many solutions have been proposed in this sector to manage routes and public transport. In our work, we will investigate the performance of sharing vehicles services.

IV. LARGE-SCALE SMART CITY TESTBED

A. OpenMTC: standard M2M Platform

The OpenMTC platform is a prototype implementation of M2M middleware, developed by Fraunhofer FOKUS and Technical University of Berlin (TUB) [15]. It has been designed to act as a horizontal convergence layer supporting multiple vertical application domains, such as transport, utilities, automotive, eHealth, etc., which may be deployed independently or as part of a common platform. The first release of OpenMTC features are aligned with ETSI M2M Rel. 1 specifications [16][17], providing an implementation of ETSI specified Service Capability Layers (SCL) at the Frontend (Gateway GSCL) and Backend (Network NSCL) M2M architecture. Currently, release 3 is finalized to support ETSI M2M Rel.2 and now the oneM2M specifications are adapted on the design of the upcoming OpenMTC release 4.

OpenMTC supports a client/server based RESTful architecture with a hierarchical resource tree defined by ETSI, and communication over all interfaces is independent of the transport protocol. The OpenMTC Reachability, Addressing and Repository (RAR) capability manages a subscription and notification mechanism. Through this mechanism, applications, gateways and the OpenMTC platform are able to receive notifications from each other, enabling management and control of devices, which belong to the same service provider or using the same technology family. As illustrated in Figure 2, the OpenMTC platform includes a Generic Transport (GT) layer that enables the interaction between the frontend and backend over unmanaged access, as well as managed access networks by integrating with the OpenEPC framework. The GT layer includes an Adaptable M2M Transport (AM2MT) module, which provides pluggable transport protocols such as Hypertext Transfer Protocol (HTTP) and Constrained Application Protocol (CoAP) [18].

OpenEPC provides enhanced communication management capacities for Quality of Service (QoS) enforcements, communication channel management and bootstrapping. The OpenMTC gateway supports a Store and Forward (SAF) feature for applications. This feature enables the handling of different traffic streams based on their priority. To support integration of heterogeneous sensors, the design of the platform enables the integration of various specific inter-working proxies. Each proxy is responsible for one technology (e.g., FS20, ZigBee, etc.) and acts as a controller for the external devices by mapping these devices for monitoring and controlling into the M2M resource tree. In order to support the development of M2M applications and make the core assets and service capabilities available to 3rd party developers, the OpenMTC application enablement consists of a set of high-level abstraction Application Programming Interfaces (APIs). These APIs hide the internal system complexity, and allow the

developer to focus on the implementation of the application logic. The system supports either XML or JSON format for data representation.

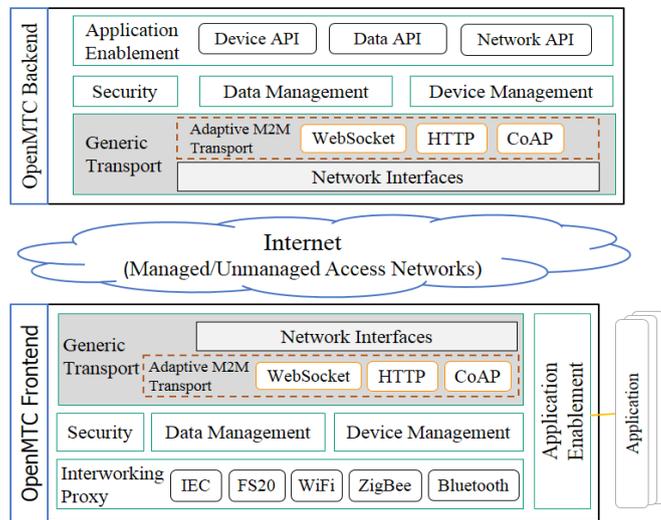


Fig. 2. OpenMTC Platform

B. FITeagle

FITeagle is a framework for managing and federating testbeds. It was developed at the TU Berlin. The intention is to use this framework for managing the interconnected testbeds at TUB and University of Cape Town (UCT). It will give the experimenter the possibility to access resources at the testbeds via the FIRE API [19].

C. Testbed deployment

The testbeds at TUB and UCT are connected via a Virtual Private Network (VPN) connection. Both shall be realized as a mix of virtual and physical components. The testbeds and their interconnection are set up in tasks of the TRESCIMO project. Both interconnected testbeds shall host several devices and gateways, for aggregating and exchanging data. The M2M Middleware core is using the OpenMTC platform and a Smart City Platform, which is developed by CSIR and is responsible for the Big Data analyses and provide APIs for experimenters, Smart City users and other Smart City Applications.

1) Overview

The testbed shall have two parts, a virtual part and a real part. Figure 3 shows the deployment at TUB and UCT. The testbed will be realized via OpenStack. That means all virtual machines are hosted inside the OpenStack Hypervisor to represent different M2M entities (e.g., devices and gateways). OpenStack also virtualizes the used networks, and provides interfaces for FITeagle allowing for readily available computing processing power for potential high loads in stress testing. The combination of OpenStack and FITeagle will empower researchers to conduct distributed experiments and verifications. Additionally, a real network shall be added in order to enable the integration of real devices/sensors into the system.

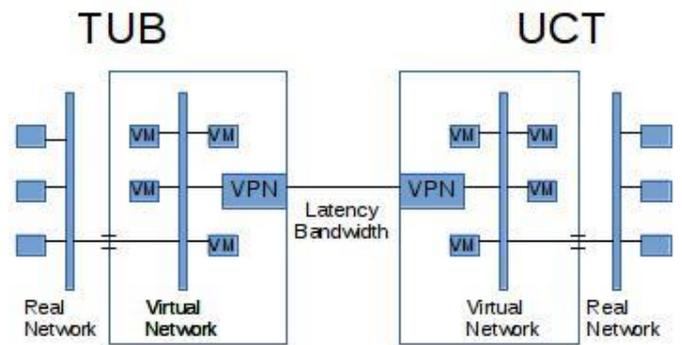


Fig. 3. large-scale M2M testbed

The connection between both testbeds shall be realized with a VPN connection, which is managed by OpenVPN, an open source program for handling of VPN connections. Because the testbeds are located at different continents, the connection will be subjected to different problems and limitations. The connection itself may be a problem due to the reliability of sea cables. Latency is certainly a problem because of the great distance and number of routers between the two testbeds and the limitation due to the speed of light. However, this will leverage the realistic behaviour of the obtained results.

2) Components

For setting up the testbed existing software was used. OpenStack was used to establish the virtual environment, the virtual machines and the virtual networks. For securely connecting both testbeds OpenVPN was used. With that, the two private networks of the testbeds are reachable from both sides.

For integrating the Smart City Platform and the M2M Middleware into the testbed, similar approaches are used for the virtual and the physical ones. On the physical nodes the components are installed as a service which can be started and stopped via operating system methods. For example “service start openmtc-nscl” on a GNU/Debian based distribution. Configuration is achieved by altering the configuration files. The physical nodes are always powered. These services are also deployed to virtual images, which are separately created per service and are used to start VMs with. Then these services are started when booting. Configuration is done by using a metadata service of OpenStack to change configuration on boot up.

Integrating the devices is possible in diverse manner. It depends on the devices itself. Devices that can be directly connected to the network are integrated with no effort. Integrating devices which are connected via different media like Bluetooth or ZigBee need a mediator in between. On the one hand, this mediator is needed to connect these devices with the M2M platform during the experiments. On the other hand, the mediator will provide mechanisms to control such devices for experiment management.

3) Status

The setup of the testbed started recently. At TUB, the virtual testbed was setup with OpenStack and the VPN server was set as a VM in the virtual testbed. At UCT, work has started to set

up the real testbed. For this purpose, some physical nodes were connected to each other and to a physical VPN server. The VPN connection was already set up and the first tests were performed.

In the future, both testbeds shall be extended so that they have a virtual and physical part. On both sides, the devices shall be added to the system and integrated into the Smart City architecture.

V. CONCLUSIONS AND FUTURE WORK

The remarkable increase in worldwide populations moving from rural into urban areas will affect the economic growth and add more challenges of planning and managing cities. Furthermore, the connected world is extending exponentially including physical objects, computers and smartphones in a global Internet of Things (IoT). For Smart Cities to be successful, they will need to leverage the utilization of past and future generations of Information, Communications and Technologies (ICT).

The need for large-scale testbeds for Smart Cities has been recognized by industry and academia, in order to develop a candidate model for the Smart City implementation. In our work, we build a large-scale testbed for the research and experimentation of enabling technologies, standardize platforms and applications for Smart City. Our work takes into consideration the different requirements and challenges of developing M2M services in both developed and developing world.

The future work will focus on conducting the designed experimentations using the implemented testbed. The first phase of the experimentation will perform functionality testing to validate the components and mechanisms of the system. Later on, services of various domains will be tested. The testbed is extensible to higher-level components, such as data analytic tools required within smart city platforms.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 611745, as well as the South African Department of Science and Technology under financial assistance agreement DST/CON 0247/2013. Authors acknowledge the collaboration of the rest of the TRECIMO consortium partners.

REFERENCES

- [1] L. A. H. G. and J. P. José M. Hernández-Muñoz, Jesús Bernat Vercher, Luis Muñoz, José A. Galache, Mirko Presser, "Smart Cities at the Forefront of the Future Internet," *Futur. Internet*, vol. 6656/2011, no. Springer Berlin / Heidelberg, pp. 447–462, 2011.
- [2] R. De Santis, A. Fasano, N. Mignolli, and A. Villa, "Smart City: fact and fiction," no. 54536. Munich Personal RePEc Archive (MPRA), 2014.
- [3] C. Balakrishna, "Enabling Technologies for Smart City Services and Applications," in *2012 Sixth International Conference on Next Generation Mobile Applications, Services and Technologies*, 2012, pp. 223–227.
- [4] A. Elmangoush, H. Coskun, S. Wahle, and T. Magedanz, "Design aspects for a reference M2M communication platform for Smart Cities," in *2013 9th International Conference on Innovations in Information Technology (IIT)*, 2013, pp. 204–209.
- [5] J. Manyika, M. Chui, J. Bughin, R. Dobbs, P. Bisson, and A. Marrs, "Disruptive technologies: Advances that will transform life, business, and the global economy," no. May. McKinsey Global Institute, 2013.
- [6] H. Chourabi, T. Nam, S. Walker, J. R. Gil-Garcia, S. Mellouli, K. Nahon, T. a. Pardo, and H. J. Scholl, "Understanding Smart Cities: An Integrative Framework," in *45th Hawaii International Conference on System Sciences*, 2012, pp. 2289–2297.
- [7] C. Harrison and I. A. Donnelly, "A Theory of Smart Cities," in *Proceedings of the 55th Annual Meeting of the ISSS*, 2011, pp. 1–15.
- [8] L. Sanchez, V. Gutierrez, J. A. Galache, P. Sotres, J. R. Santana, and J. Casanueva, "SmartSantander: Experimentation and Service Provision in the Smart City," in *Wireless Personal Multimedia Communications (WPMC), 2013 16th International Symposium on*, 2013, pp. 1–6.
- [9] Z. Padrah and T. Solc, "Network design for the LOG-a-TEC outdoor testbed," 2nd International Workshop on Measurement-based Experimental Research, Methodology and Tools, Dublin, Ireland. 2013
- [10] T. Olivares, F. Royo, A. M. Ortiz, and I. Mines-telecom, "An Experimental Testbed for Smart Cities Applications," in *Proceedings of the 11th ACM International Symposium on Mobility Management and Wireless Access*, 2013, pp. 115–118.
- [11] "TRECIMO (Testbeds for Reliable Smart City Machine to Machine Communication) Project." [Online]. Available: www.trecimo.eu.
- [12] J. Mwangama, A. Willner, N. Ventura, A. Elmangoush, T. Pfeifer, and T. Magedanz, "Testbeds for Reliable Smart City Machine-to-Machine Communication," in *Southern African Telecommunication Networks and Applications Conference (SATNAC)*, 2013.
- [13] A. Elmangoush, Andreea Corici, Marisa Catalan, R. Steinke, Thomas Magedanz, Joaquim Oller "Interconnecting Standard M2M Platforms to Delay Tolerant Networks," in *The 2nd International Conference on Future Internet of Things and Cloud (FiCloud-2014)*, 2014.
- [14] oneM2M, "OneM2M." [Online]. Available: <http://onem2m.org/>.
- [15] "OpenMTC platform." [Online]. Available: <http://www.open-mtc.org/index.html>.
- [16] ETSI TS 102 690 v1.1.1, "Machine-to-Machine communications (M2M); Functional architecture," 2011.
- [17] ETSI TS 102 921 v1.1.1, "Machine-to-Machine communications (M2M); mla, dla and mld interfaces," 2012.
- [18] Z. Shelby, K. Hartke, and C. Bormann, "RFC 7252: The Constrained Application Protocol (CoAP)." p. 112, 2014.
- [19] "FITEagle- Future Internet Testbed Experimentation and Management Framework." [Online]. Available: <http://fiteagle.org/>.

ICT-enabled solutions for smart management of water supply in Africa

P.J.C. Nel*, M.J. Booysen*, B. van der Merwe[†]

*Department of Electrical and Electronic Engineering and MTN Mobile Intelligence Laboratory, Stellenbosch University, Stellenbosch, South Africa.

[†]Department of Computer Science and MTN Mobile Intelligence Laboratory, Stellenbosch University, Stellenbosch, South Africa.

Email: {15634280, mjbooyesen, abvdm}@sun.ac.za

Abstract—Pervasive and ubiquitous technologies that include mobile device applications, machine to machine communications, and cloud computing, are increasingly used for cost-effective data aggregation and information dissemination. Recently, this trend has started to gain momentum in the water sector and is being used for various management and monitoring tasks, such as remote leakage detection, automated meter reading and enhanced usage feedback to water users. This paper analyses the challenges faced by various stakeholders (consumers, utilities, etc.) in the water supply industry. Application of the said technologies is then proposed to address these unique challenges and the varying data needs of all stakeholders. An example solution, with a mobile device application and supporting cloud computing solution, is developed and presented as a proof-of-concept to further illustrate the potential use of ICT for water supply management.

I. INTRODUCTION

Access to a safe and reliable water supply impacts every aspect of our lives, even if we may not be aware of the consequences. For over 300 million people living in Africa without access to safe water [1], hours are spent every day fetching water from rivers and other potentially unsafe sources of water simply to meet their daily water needs. Additionally, they face the risk of infection from water borne diseases or dehydration from using unsafe and vulnerable water sources.

For those with access to piped water in their homes, it is easy to take this source of water for granted and forget about the water scarcity being experienced around the world. The only usage feedback they receive is a monthly bill that details the user’s cumulative consumption at the end of a month, making it difficult to identify the effects of specific events (e.g. how much energy and water a five minute shower uses). Additionally, many water service providers (WSPs) are struggling to recover costs associated with water service provision and the development of infrastructure. For example, in 2010, the World Bank estimated that US\$ 17 billion would be required annually for the water supply infrastructure in sub-Saharan Africa to meet the millennium development goals, with around a third of this funding allocated to operation and maintenance costs [2].

By providing consumers and service providers with additional data and enhanced means of information collection, through the use of information and communication technologies (ICTs), many of these challenges can be alleviated. These technologies have been leveraged in other sectors, such as

healthcare, education and agriculture, to create unique solutions for the problems faced by these industries [3]. For example, Text to Change (TTC) has utilised basic mobile phones to provide farmers in Africa with daily information on effective agricultural practices, market prices and weather forecasts using text and voice messages [4].

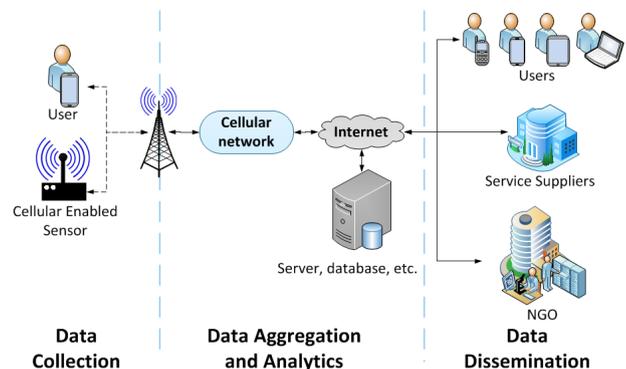


Fig. 1. Typical flow of information in the water supply industry. (Adapted from [5])

Figure 1 shows the typical flow of information for ICT systems in the water supply industry. Data collection can be performed by users (customers or employees of WSPs) with a mobile device through various means, including Short Message Service (SMS), Unstructured Supplementary Service Data (USSD) or data submissions from smartphone devices (e.g. photos or co-ordinates). Alternatively, information can be reported in by automated sensors over cellular networks. The data is then stored in a central server or database where it can be accessed directly or processed further to create useable data (e.g. reports for WSP managers). Data can then be accessed by users (e.g. consumers accessing usage data) or disseminated automatically to relevant stakeholders (e.g. informing users of events).

A. Contribution

This paper presents an analysis of the various implementations of ICTs in the water supply industry in Africa and illustrates how these technologies can be used to empower a wide array of users, with varying economic and physical (i.e. location) circumstances, as well as offering benefits to all the stakeholders in the industry. Additionally, a proof-of-concept

Android mobile application is also presented, which can be used to view usage data for domestic hot water cylinders.

The rest of this paper is organised as follows: Section II describes the various roles that ICTs can play in the water industry and the challenges they can help to alleviate; Section III presents the proof-of-concept Android mobile application that is used to further illustrate the potential of ICTs as well as the future work for the application; and Section IV concludes the paper.

II. ROLE OF ICTS IN THE WATER INDUSTRY

The mobile industry has had a tremendous effect on the African continent. Global System for Mobile Communications (GSM) coverage is estimated to reach 76 percent of the African population and has outgrown access to reliable and affordable electricity and water services over the past 10 years [6]. The rapid growth of mobile network coverage is providing millions of people with first time access to modern infrastructure services. For example, 130 million people in sub-Saharan Africa (SSA) are covered by mobile networks but do not have access to an improved water source [6]. Mobile phones are becoming increasingly ubiquitous, with a unique mobile subscriber penetration of 35 percent in Africa at the end of 2013. Along with the decreasing cost of mobile handsets, these devices are being used as cost-effective tools for collecting and disseminating critical information [5] [7]. **Mention penetration and growth rates of smartphones in Africa.

Coupled with the growth in machine-to-machine (M2M) communications, mobile banking and other ICTs, there are many opportunities to enhance the lives of everyone in Africa. The following sections present examples of using ICTs for various stakeholders and applications in the water industry in Africa, ranging from cellular enabled handpumps to increase reliable service provision for the rural poor, to smart meter enabled systems to aid middle- and high-income groups in urban areas in using their energy and water more efficiently.

A. Crowdsourcing Data

Water quality is a problem that affects all stakeholders in water supply industry: WSPs are required to monitor the quality of the water they provide; and end users face the health risks associated with consuming contaminated water. In crowdsourcing systems, data is collected and submitted by members of the community via mobile devices [8]. The aim of these systems is to allow end-users to directly communicate information to relevant stakeholders in order to benefit both parties [5]. For example, the Mobile 4 Water application allows members of the community, in Uganda, to report a fault at a water point via SMS [9]. Over 15 000 water points are being monitored by this system, which allows WSPs to obtain additional information which would not have been possible through costly field visits [9]. End-users, on the other hand, are incentivised with the promise of better service provision for communities by prompting WSPs to take corrective action [8], as well as being able to view the status of water points in their area and, therefore, avoid consuming water from contaminated sources [9].

Another example of such as system is the m-Maji mobile phone-based water information system in Kenya. Water vendors, who resell water from the utility network or a private source, play a crucial part of service provision in many African countries [10]. For example, mobile vendors serve an estimated 32 percent of the urban population in Mauritania [11]. This is because private piped water supply access is limited and, in areas where it is available, service provision is often intermittent or insufficient [10] [11] [12]. The m-Maji mobile application allows water vendors to advertise their services using USSD. This includes: the selling price of their water; their location; and, optionally, whether or not the water has been purified. This data is stored in a central database which water buyers can use to search for suitable vendors, also using USSD. The system also caters for quality management by allowing buyers to report vendors who are fraudulent (e.g. lied about quality or price) to alert future buyers [13]. Additionally, the quality of water being advertised as purified is monitored by M-Maji staff through random monthly water quality tests [13].

B. Data Capturing

In addition to using crowdsourcing as a means of collecting data, mobile phones can also be used to facilitate the process of data capture by WSPs' field staff. In [14] the design of a mobile phone-based information system is presented for water quality management by municipalities in rural and under-resourced areas. The system was implemented in four rural municipalities in South Africa and employed two mobile applications: the water quality reporter (WQR); and the water quality manager (WQM). The WQR application is used for data collection and was installed on the phone of a water supply caretaker (e.g. borehole operator). The caretaker could collect and insert the required data into a form that is then submitted to a server for verification and storage purposes. The server would return a message that informs the caretaker of the success or failure of the submission. In the event that the caretaker submits any erroneous data (i.e. invalid values), the information would have to be recaptured and resubmitted.

Additionally, a message is sent to the responsible manager if the results are outside of suitable bounds. The WQM application is intended to support management functions, such as analysis of collected data, and was thus installed on four Android phones which were supplied to a manager at each municipality. Managers could review water quality testing for any of the water sources under their authority in real-time. The WQR system automatically provided the relevant managers with weekly reports of raw data in spreadsheets. The reporting of water quality standards of all four municipalities increased with the introduction of this system. Additionally, managers and water supply caretakers stated an increased awareness and appreciation for collecting data used in monitoring and diagnostics.

C. Remote Asset Monitoring

In Africa, handpumps are installed in remote rural locations as a means of improving the access to safe drinking water and to decrease the distance the rural population must travel to satisfy their water needs. The operation and maintenance (e.g. spare parts for repairs) of these assets is a complex challenge

[15] as there is insufficient data on their usage and status. Often the spare parts required for repairs are not always rapidly available [15] and it has been estimated that a third of installed handpumps are out of service at any given time [16]. GSM coverage is expanding into areas that rely on handpumps for their water supply, allowing ICTs to facilitate the management of these assets [8].

The design, building and testing of a prototype Waterpoint Data Transmitter (WDT) that provides real-time usage data for handpumps is presented in [15]. A field test was conducted on three India Mark 2 handpumps (installed two years prior to the test) over a period of four days in Valley View, north-west Lusaka (Zambia). The WDT is attached to the handle of a handpump and consists of: a simple microprocessor; an integrated circuit (IC) based accelerometer; and a Global Systems for Mobile communications (GSM) modem. The device monitors the number of strokes made during operation of the handpump and transmits this data in one minute intervals via SMS using the GSM network. Analysis of this data can lead to: estimates of the usage (volumetric output) and performance of handpumps; the detection of faults and possibly insight into the nature of the failure when compared to historical data [15]; better infrastructure management by identifying areas requiring additional assets; and increased accountability of service suppliers, as the upkeep and usage of assets can be recorded to determine the efficiency of service delivery. Additionally, all of these functions can be performed remotely which implies that less field visits are required to collect the relevant data (e.g. determining usage of a specific asset), resulting in substantial cost reductions for WSPs.

D. Information Dissemination

Another challenge faced in Africa is water-borne diseases, which kill millions of people each year. This section presents the mHealth E.coli smartphone application. This application is used in conjunction with a Mobile Water Kit (MWK) to act as a low-cost water monitoring system for the rapid detection of certain harmful bacteria (i.e. E.coli and total coliform) in water samples [17]. To assess the quality of water, the user is required to collect a sample of water to be tested using a syringe and filtering the sample through a syringe filter unit. The user must then add the formulated chemical reagents sequentially onto the syringe filter unit. The presence of harmful bacteria is indicated by the appearance of a red colour on the surface of the syringe filter unit.

This data can then be submitted via the mHealth E.coli application, which runs on an Android smartphone. Users first select the type of source that the water sample was collected from (e.g. river, well, etc.) and then proceed to take a photo of the tested syringe filter unit. The photo is analysed by the application and the results are shown to the user. Thereafter, the photo and the Global Positioning System (GPS) location of the smartphone are uploaded to a server via SMS to provide a unique location identifier for contaminated water sources. These results are then used to send SMS notifications to issue water quality alerts to subscribed users. Although this system was trial-tested in Canada and India, it is also applicable to the African context. This system can create an early warning system to detect outbreak of water-borne diseases in communities and users of contaminated sources.

E. Consumer Data Access

Eco-feedback is based on the theory that the majority of users are unaware of or don't understand the impact of their daily activities on the environment [18]. This lack of knowledge can be mitigated using ICTs to help consumers better understand their usage by accessing higher resolution data through timeous feedback channels. In [19], a wireless M2M network was used in several regions of South Africa to remotely monitor and control the hot water cylinders (HWCs) of residential households in near real-time. The HWCs are equipped with sensors (e.g. thermocouple) and actuators (e.g. dump valve) that are used to collect data and perform control operations, such as monitoring temperature or safely emptying the HWC in case of a failure. A cellular modem allows these devices to connect to a remote server with general packet radio service (GPRS) using cellular connections. Data can then be reported to the server at a set interval and is stored in a database where it can be used in its raw form (e.g. current temperature of water in HWC) or processed further to create useful results (e.g. cumulative energy usage for a day). Users can access this stored data via an Internet portal which performs queries on the database and relays relevant data to corresponding users and also allows them to issue control commands, such as manually turning a HWC on or off. Commands are received by the assets from their server using the cellular modems and are performed by the hardware (e.g. microprocessor and actuators) situated at the HWC asset. This system therefore allows consumers to view their usage, such as temperature and energy graphs, and control aspects of the system from any remote location with Internet access.

Feedback of usage data to consumers in near real-time can allow them to make informed decisions regarding their future resource use, resulting in reductions in water and energy demand from residential households [20]. This type of system is better suited to middle- and high-income users as it requires users to live in a household with electricity and private piped water access, as well as requiring users to have Internet access to interact with the systems online interface. However, this system illustrates how ICTs can be leveraged to create a wide array of solutions for all types of consumers. Additionally, this type of system is beneficial to WSPs as they can better understand consumers usage (on a local, provincial or national level) and improve management of infrastructure and other management tasks with higher resolution data [20]. For example, they could use a smart metering system to determine if users are complying with water restrictions, which is of particular importance to drought stricken areas. This data can further support management tasks for WSPs by allowing them to cater education programs (aimed at promoting water conservation) towards specific usage patterns in different areas (e.g. a program could focus on a specific end-use of water, such as outdoor use for gardening activities).

III. PROOF-OF-CONCEPT SMARTPHONE APPLICATION

The interface used to control assets, mentioned in the previous section, is not particularly well suited for access by mobile phones or tablets. The proof-of-concept mobile application presented in this section, therefore, builds upon the functionality of the aforementioned system by allowing users to monitor their usage through an Android smartphone

by interacting with easy to use native Android elements, such as spinners (i.e. dropdown menus).

This section presents a proof-of-concept smartphone application that is aimed at aiding demand-side management of residential household energy and water usage (i.e. reduction in water demand of residential consumers), which, in turn, allows for the deferral of additional water supply infrastructure asset construction [20] and the conservation of precious water resources. Although HWCs make use of both water and electricity and only includes hot water usage (not total household water usage), the application constitutes as a eco-feedback implementation. This mobile application runs on an Android smartphone and allows users to monitor the current status of their HWC as well as graph historical usage data. Data is requested from the central server through the system's representational state transfer (REST) application programming interface (API) and suitable Hypertext Transfer Protocol (HTTP) GET requests, based on the task the user is performing. The server can either convey the raw data to users (e.g. instantaneous thermostat temperature) for graphing purposes or it can provide users with metric values, which have been processed by the server to create a useful data value (e.g. average internal HWC pressure) for asset status updates or monitoring purposes. The application is divided into three main tabs that are described in further detail in the following sections.

A. Usage Tab

The Usage tab, shown on the right in Figure 2, can be used to view the most recent data reported by the HWC asset. It provides the user with an overview of the asset's daily resource consumption and the associated cost thereof (in Rands). The data is also timestamped so that users are able to determine when last these values were updated. Users are also able to manually refresh this data by clicking the refresh icon in the action bar. If the application attempts to obtain data from the server and is unsuccessful (due to lack of network access, etc.) a relevant error message is displayed to the user. The system is not yet able to provide usage statistics for user water consumption and thus the values shown for water consumption and flow are included to provide a preview of future work and to illustrate the complete appearance of the user interface.

The data summarised by this tab can help users to better understand their daily usage patterns without providing excessive amounts of data that would overwhelm and confuse users. Additionally, the tab provides both a usage value (either kWh or litres), for those users who are perhaps more technical or interested on the environmental impact of their usage, as well as a monetary value, for users who may be non-technical users or interested in financial incentives for reducing their energy and water consumption.

B. Status Tab

Users are able to view the status and present settings for the control unit that manages the HWC asset on the Status tab shown on the left in Figure 2. This includes: the present water temperature value; the temperature set point, which is only enforced when the control unit is enabled; the present on or off state of the HWC element; the present energy usage

of the asset; the status of the timer control unit (i.e. whether or not control unit is enabled); and the on/off schedule that is implemented by the control unit (which is an expandable table that can be minimised to save space). However, due to API restrictions, users must login separately on the Internet platform, through a browser window, to edit any of the settings of the controller unit (such as the schedule or temperature set point). The dials used in the usage and status tabs were created by editing a custom library, called SpeedometerView [21].

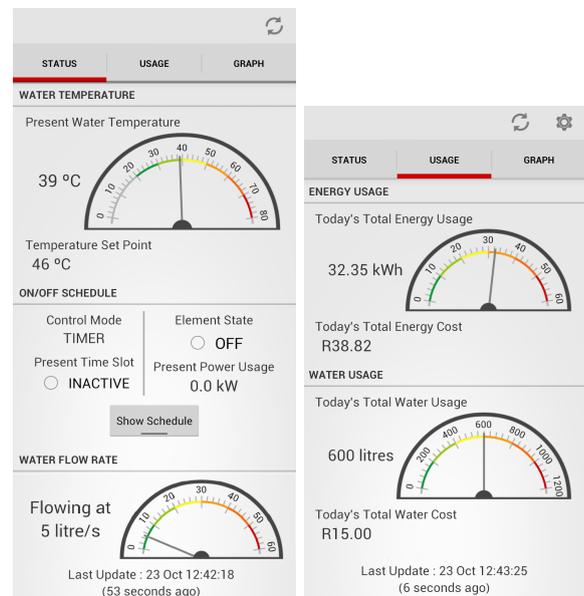


Fig. 2. Status (Left) and Usage (Right) tab of smartphone application

C. Graph Tab

The application also allows users to graph various types of daily usage data on the Graph Tab, shown in Figure 3, in order to better understand their usage. This functionality was created with the use of AChartEngine, an open source graphing library for Android [22]. The user can choose the type (e.g. energy consumption) and date of the data they wish to view. The data for an entire day is obtained from the server and displayed in time intervals of one minute. Users are also able touch anywhere on the graph to obtain the time and data values for a particular point (the annotation and vertical line in Figure 3).

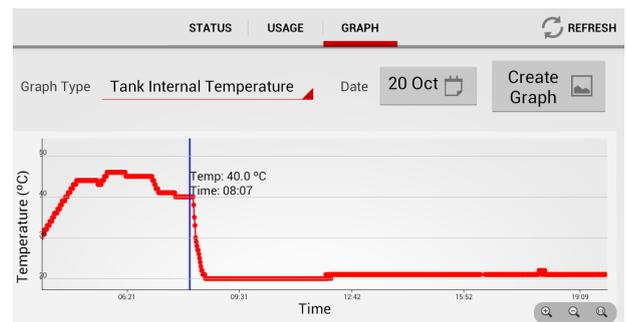


Fig. 3. Graph tab of smartphone application

D. Future Work

Future development of the mobile application will involve several enhancements to the functionality. Firstly, the application will include the feedback of water usage to consumers as this is important in understanding their overall resource consumption. Secondly, control capabilities will be added to the application to allow users to control the various aspects of their HWC (such as the temperature set point and on/off schedule) from the mobile application interface. Additionally, the analytics of the system will be greatly improved. The system will access users' current settings and analyse their daily usage patterns to make recommendations on optimising their HWC's settings in order to reduce their energy usage. Coupled with the control functionality, this has the potential to significantly reduce consumer usage by disabling the HWC when not in use while maintaining the temperature of the user's hot water at a comfortable level that suits their schedule. Furthermore, an audit function will allow users to enter information about their household (e.g. the number of bathrooms, showers, etc.). This information will then provide users with a typical usage value that would be expected for such a household. Users could then compare their usage to these values (i.e. normative comparisons) and become aware of how much water they use in comparison to other similar households, which helps to contextualise a user's consumption by providing a benchmark for resource usage.

IV. CONCLUSION

This paper presented several ICT-based solutions to challenges faced by the water supply industry in Africa. The proposed solutions cater for a variety of stakeholders with varying information needs and circumstances, ranging from low-income water consumers in remote rural areas to WSPs serving middle- to high-income consumers in an urban setting. A proof-of-concept smartphone application was also presented to illustrate how these technologies can be leveraged to create unique solutions. This mobile application was aimed at increasing consumer awareness and reducing water and electricity demand, as well as providing WSPs with additional data for management tasks.

ACKNOWLEDGMENT

The authors would like to acknowledge: Trinity Telecoms for their technical support and guidance, and providing access to the SMART M2M-enabling system; and MTN for their continued support and funding.

REFERENCES

- [1] WHO & UNICEF, "Progress on Sanitation and Drinking-water 2013 Update," 2013. [Online]. Available: http://www.who.int/water_sanitation_health/publications/2013/jmp_report/en/
- [2] Vivien Foster and Cecilia Briceno-Garmerñia, Ed., *Africa's Water and Sanitation Infrastructure: A Time for Transformation*, 2010, ch. 16 - Water Supply: Hitting the Target?, DOI: 10.1596/978-0-8213-8041-3.
- [3] GSMA, "Sub-Saharan Africa Mobile Economy 2013," Tech. Rep., November 2013. [Online]. Available: http://www.gsmamobileeconomyafrica.com/Sub-Saharan%20Africa_ME_Report_English_2013.pdf
- [4] A. Swank. (2013, March) mAgr - Developing mAgriculture Systems: how does Text to Change do it? [Online]. Available: <http://www.gsma.com/mobilefordevelopment/developing-magriculture-systems-how-does-text-to-change-do-it>
- [5] M. T. Hutchings, A. Dev, M. Palaniappan, V. Srinivasan, N. Ramanathan, and J. Taylor, "mWASH: Mobile Phone Applications for the Water, Sanitation, and Hygiene Sector," Pacific Institute & Nexleaf Analytics, Tech. Rep., April 2012. [Online]. Available: <http://pacinst.org/wp-content/uploads/sites/21/2014/04/mwash.pdf>
- [6] M. Nique and K. Opala, "The Synergies Between Mobile Energy and Water Access: Africa," GSMA, Tech. Rep., March 2014. [Online]. Available: http://www.gsma.com/mobilefordevelopment/wp-content/uploads/2014/04/MECS_Synergies-between-Mobile-Energy-and-Water-Access_Africa.pdf
- [7] GSMA, "GSMA Mobile Enabled Community Services - Annual Report 2013," Tech. Rep., May 2014. [Online]. Available: http://www.gsma.com/mobilefordevelopment/wp-content/uploads/2014/05/GSMA_MECS_Annual-Report-2013.pdf
- [8] P. Thompson, R. A. Hope, and T. Foster, "Is silence golden? of mobiles, monitoring, and rural water supplies," *Waterlines*, vol. 31, no. 4, pp. 280–292, October 2012, DOI: 10.3362/1756-3488.2012.031.
- [9] (N.d) About Mobile 4 Water. [Online]. Available: <http://m4water.org/pages/aboutus.php>
- [10] S. Banerjee, H. Skilling, V. Foster, C. Briceño Garmendia, E. Morella, and T. Chfadi, "AICD Background Paper 12 Ebbing Water, Surging Deficits : Urban Water Supply in Sub-Saharan Africa," World Bank, Tech. Rep., June 2008, DOI: 10986/7835.
- [11] S. Keener, M. Luengo, and S. Banerjee, "Provision Of Water To The Poor In Africa : Experience With Water Standposts And The Informal Water Sector," World Bank, Tech. Rep., August 2010, DOI: 10.1596/1813-9450-5387.
- [12] S. Banerjee, Q. Wodon, A. Diallo, T. Pushak, H. Uddin, C. Tsimpo, and V. Foster, "AICD Background Paper 2 Access, Affordability and Alternatives: Modern Infrastructure Services in Africa," World Bank, Tech. Rep., February 2008, DOI: 10.1596/978-0-8213-8457-2.
- [13] (N.d) How M-Maji Works. [Online]. Available: <http://mmaji.wordpress.com/m-maji/>
- [14] U. Rivett, M. Champanis, and T. Wilson-Jones, "Monitoring drinking water quality in South Africa: Designing information systems for local needs," *Water SA*, vol. 39, no. 3, pp. 409–414, 2013, DOI: 10.4314/wsa.v39i3.10.
- [15] P. Thompson, R. A. Hope, and T. Foster, "GSM-enabled remote monitoring of rural handpumps: A proof-of-concept study," *Journal of Hydroinformatics*, vol. 14, no. 4, pp. 829–839, 2012, DOI: 10.2166/hydro.2012.183.
- [16] RWSN Executive Steering Committee, "Myths of the Rural Water Supply Sector: RWSN Perspectives No. 4." RWSN, St Gallen, Switzerland, Tech. Rep., 2010.
- [17] N. S. K. Gunda, S. Naicker, S. Shinde, S. Kimbahunu, S. Shrivastava, and S. Mitra, "Mobile Water Kit (MWK): a smartphone compatible low-cost water monitoring system for rapid detection of total coliform and E. coli," *Analytical Methods*, vol. 6, no. 16, pp. 6139 – 6590, August 2014, DOI: 10.1039/c4ay01245c.
- [18] J. Froehlich, L. Findlater, and J. Landay, "The design of eco-feedback technology," in *ACM Conference on Human Factors in Computing Systems*. ACM, 2010, pp. 1999–2008, DOI: 10.1145/1753326.1753629.
- [19] M. J. Booysen, A. Molinaro, and J. A. A. Engelbrecht, "Proof of Concept: Large-Scale Monitor and Control of Household Water Heating in Near Real-Time," in *International Conference on Applied Energy ICAE 2013*, 2013, DOI: 10019.1/85478.
- [20] D. P. Giurco, S. B. White, and R. A. Stewart, "Smart Metering and Water End-Use Data: Conservation Benefits and Privacy Risks," *Analytical Methods*, vol. 2, no. 3, pp. 461 – 467, August 2010, 10.3390/w2030461.
- [21] A. Danshin. (2014, February) SpeedometerView Version 1.0.1. [Online]. Available: <https://github.com/ntoskml/SpeedometerView>
- [22] (2014, August) AChartEngine Version 1.2.0. [Online]. Available: <http://www.achartengine.org/>

Design and Development of a Satellite Based Water Resources Monitoring System

Harry Mafukidze

Department of Electrical and
Electronic Engineering
University of Stellenbosch
Email: 16736893@sun.ac.za

Riian Wolhuter

Department of Electrical and
Electronic Engineering
University of Stellenbosch
Email: wolhuter@sun.ac.za

Abstract—The Faculty of Forestry and Woodscience at Stellenbosch University has a requirement to monitor and record water resources and environmental data at remote sites, not within reach of any mobile services. The current solution consists of a standalone data logger based monitoring system. This system, however, is not ideal as it does not provide data in real time and has high costs and other problems servicing the particular sites. This paper presents an alternative satellite based WSN (Wireless Sensor Network) solution to this problem. The system described in this paper comprises a WSN with a three-part framework. The first part consists of sensor nodes monitoring rainfall, air temperature, air humidity, ambient light, wind speed, wind direction, soil temperature and soil moisture. Communication from these nodes to the central gateway is based in the wireless ISM band. The second part contains an Iridium satellite communications module, a gateway with a Linux based SBC (Single Board Computer) for collecting, storing and sending data from sensor nodes and forwarding such data via the SBD (Short Burst Data) satellite messaging service. The third part consists of the MS (Master Station), which is used for displaying sensory and site information. The system is solar powered and measurements indicate that the system meets an overall standby time of at least three days, as stated in the project requirements. It has been tested continuously in an actual deployment situation and is performing well. This new satellite based monitoring system is certainly an improvement and a reliable alternative to the one used up to now.

I. INTRODUCTION

Efficient water resources monitoring is a major concern for the Department of Forestry. South Africa has a semi-arid climate and is a relatively dry country with a mean annual rainfall of 480mm. Of that value, only 9% is converted to river runoff [1]. The Western Cape, however, has a Mediterranean type climate which receives its rain in winter. As a result of climate change, available water resources, runoff and ground-water resources will be affected [1]. Information on rainfall, air temperature, air humidity, ambient light, wind speed, wind direction, soil temperature and soil moisture is of paramount importance to researchers, hence the need to put in place monitoring mechanisms. With the need for remote monitoring, focus has turned to satellite based monitoring methods. Taking this as the motivation of our project, we use a satellite link to transfer sensory information from the GS (Ground Station) to the MS. Features of our system include a Zigbee based WSN to send data to the GS. The WSN consists of four different sensor nodes which monitor site-specific data mentioned above. The GS comprises of a Linux based SBC which stores data and

communicates with the IR modem. The MS is a web based application which displays the sensory data.

II. BACKGROUND

Previous work on monitoring water resources has been discussed in a number of articles and some of them include systems which employ GSM and blue-tooth systems. CEDEC [2] developed a fresh water real-time monitoring system based on WSN and GSM. The system uses RF XBEE 802.15.4 which operates in the 2.4GHz spectrum, PIC16F886 MCU and the coordinator device is interfaced with a GSM/GPRS modem. [3] developed a Virtual Instrument for Radio Telemetry. In their work, they used a GP300 transceiver from Motorola to transmit environmental conditions to a base station. The authors in [11] proposed a WSN for temperature monitoring. They developed the system to manage air conditioning systems at their institute. In their work, they developed a system which consisted of three main blocks: data acquisition, data collection and data display. Similarly, our system incorporates these blocks. As need for infrastructure-less remote monitoring increases, [4] developed an Oceanic drifter based on Iridium. In his work, drifting buoys provide information about surface currents, position and sea temperature. To provide power, [9] examined the possibilities of using solar energy to power the sensor networks. [10] implemented and utilised solar power in their WSN and found the solution to be feasible and reliable. Likewise, our project will harness power from the sun. The objective of this research is to design and develop a dedicated network to monitor water resources in any remote area and deliver data in near real-time.

Previous solutions included setting up USB/Micro SD data loggers. The data would be collected manually after some time. Ideally, an automatic system which works in any remote area without the need for existing telecommunication infrastructure is required and this system should send data automatically to the master station (MS) without human intervention. This will help reduce the costs of site visits and it also covers areas which are not serviced by current GPRS/GSM techniques.

After introducing related work in monitoring water resources, this paper defines the requirements of the specific system in section III. Section IV presents the methodology. Section V discusses the project implementation and section VI presents the results of the experiments conducted. Finally, conclusions and future work are discussed in section VII.

III. REQUIREMENTS

As discussed in the previous section, a proposed solution should be capable of remote placement, measure a number of parameters reliably and be self sustaining energy wise for worst case winter scenarios. At least three days of autonomy is required. It should also be flexible in deployment to add additional measurement nodes. The network should manage disappearance or addition of nodes without any problem. The design is affected by hardware constraints such as low power, and the components of each node should fit in a small to medium IP65 rated plastic enclosure. The size and the architecture of the network are determined by the sensing area and the parameters to be monitored.

IV. METHODOLOGY

To address the challenge at hand, available open-hardware and open-software tools for development were sought. The hardware/software platform selected for the sensor nodes was the Arduino platform [6] primarily due to its cost, availability, lower power consumption and its small form factor.

The XBee (Zigbee 2.4GHz, Series 2) modules were selected for setting up radio links and they are integrated by the Arduino hardware together with one or more sensors to form a sensor node. The ground station hosts one radio configured as a coordinator, Raspberry Pi SBC running LINUX OS which enable it to execute programs simultaneously and finally, the GS incorporates an Iridium modem to set up satellite links with the Iridium gateway. The whole system includes four different sensor nodes which monitor wind speed, wind direction, air temperature, air humidity, soil temperature, soil moisture, rainfall and ambient light.

By utilising data from these sensors over time, important parameters such as soil absorption and drainage rates could be determined. This information is used in water resources mapping and rainfall run-off studies.

The architecture and the components selected for this project was based on the requirements and constraints of the system laid out in section III.

TABLE I. COMPARISON OF RADIO POWER CONSUMPTION

	Rx current (mA)	Tx current (mA)	Link budget (dBm)	Output power / (frequency Band) (dBm)
Xbee	40	40	110	3 / 2.4GHz
XBee Pro - S2	45	295	119	18 / 2.4GHz
AT86RF230	16.0	17.0	104	3 / 2.4GHz
AT86RF212	9.0	18.0	120	10 / 700/800/900 MHz
MC13192	42	35		
CC2420	18.8	17.4		4 / 2.4GHz

Table I displays some of the available RF modules which could be used. Despite a higher current consumption, we chose the Series 2 Digi XBee radio due to its superior capabilities. Some of them include its ease of use, greater range, its ability to automatically form self-healing mesh networks and the module operates in the free unlicensed spectrum band (2.4GHz).

We selected off the shelf hardware components, shown in Table II and customised them to suit our project. The table also shows the hardware selected for sensor nodes and the GS.

TABLE II. HARDWARE REQUIREMENTS

Component	Product	Description	Specifications
Sensor Node	Arduino Uno	Arduino MCU Board	
	LEA-4P-T-MRT	GPS Module	
	Davis 6470	Soil Temp sensor, resistive	-40 ⁰ – 65 ⁰ C
	Davis 6440	Soil moist sensor, resistive	0 – 200 cb
	Davis 6382	Air Temp/humid,SH1x sensor	1 – 100% RH
	Davis 7852	Rain gauge, digital	0.2mm
Ground Station	Davis 6410	wind sensor, digital	1 – 322 km/h
	XB24-Z7WIT-004	Xbee Radio	2.4GHz, 120m (LOS)
	Raspberry Pi	SBC	Linux OS
	XB24-Z7WIT-004	Xbee Radio	2.4GHz, 120m (LOS)
Master Station	IR	5.5 VDC Iridium modem	-50 ⁰ – 85 ⁰ C
		Display sensor data	

The selected sensors shown in table II above cover wide ranges of the sense parameters and are therefore suitable for our application.

V. IMPLEMENTATION

The system is divided into two main parts which are data acquisition and data retrieval. The data acquisition covers the sensor node and the sink node whereas data retrieval covers the Master station. The project layout is shown in fig 1 below.

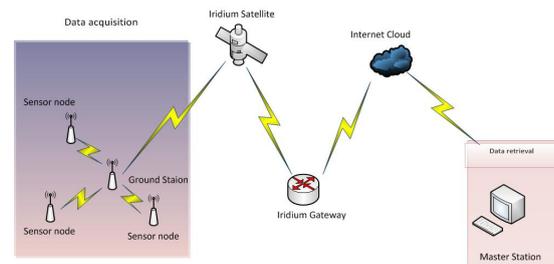


Fig. 1. Project layout.

As per Fig 1, the Zigbee network adopts a star topology. The coordinator forms the network and other routers are able to join. Data is sent to the coordinator and the coordinator forwards the data to the Iridium system. Data is then sent from the Iridium ground station to our client computer which provides a web user interface to the system.

A. Data acquisition

1) *Sensor node*: In total, four sensor nodes were built and each of them monitored a different environmental parameter. Each sensor node is equipped with 5 individual units as shown in fig 2. They are the sensing unit, processing unit, transceiver unit, power unit and the GPS unit [5]. A sensor node can communicate directly with the GS if it lies within each other's transmission range, or they can communicate through intermediate nodes.

The sensor samples the environment and generates an analog signal which is converted to a digital signal by the ADC. The micro-controller or processing unit processes the data, performs calculations and sends the data to the transceiver for transmission over a wireless link.

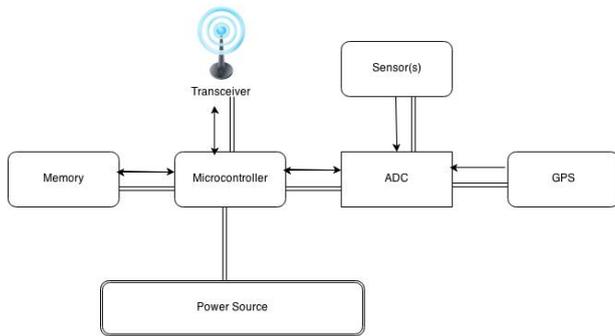


Fig. 2. Sensor node architecture

2) *Arduino* : Arduino [6] refers to a family of hardware which uses free, open source and cross platform software. An Arduino Uno, Fig 3 is based on the Atmega 328p [7], 8 bit microcontroller which clocks at 16MHz. It contains 128kB of program memory and 4kB of SRAM. It has 28 pins of which 14 can be used for digital applications and 6 can be used as PWM outputs and 6 ADCs can be used for analog applications. It features a variety of interfaces, among them 1 USB port, 1 UART TTL (5V), I2C, power jack, ICSP header.



Fig. 3. Sensor node.



Fig. 4. Arduino IDE.

The setup of the sensor node is shown in Fig 3 above and it allows us to add custom hardware and to easily control power to the radio and sensors. The Arduino IDE, shown in Fig 4, allows the developer to program the Arduino hardware.

B. Ground Station

The GS (sink node) is composed of a Raspberry Pi SBC, XBee coordinator radio and an IR modem. The SBC has a GPIO which provides UART port and a USB port for connectivity with the modem.

1) *Raspberry Pi*: The Raspberry Pi [8], shown in Fig 5, is a small sized single board computer running Linux OS.

C. Data retrieval

The MS includes a PHP web page for data viewing and analysis. Data can be manually uploaded to the server or automatically by setting the mail settings. Fig 6 shows the architecture of the database. The diagram displays information about the system user, sensor information and ground station



Fig. 5. Raspberry Pi.

information

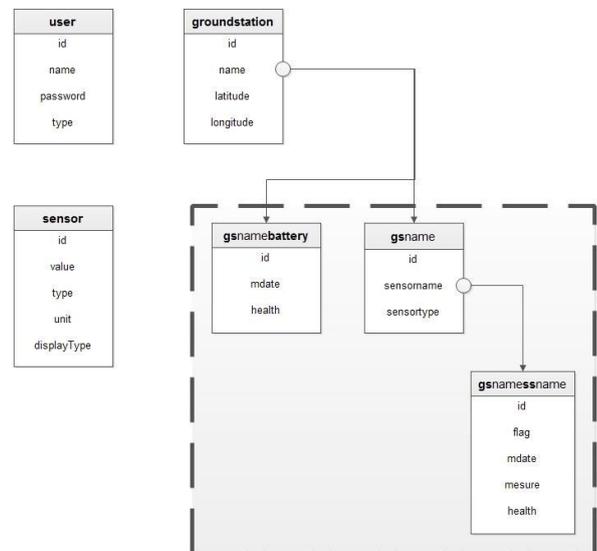


Fig. 6. Master station database

D. Software

The software for the sensor nodes is developed with the Arduino platform [6] and is shown in Fig 4. The code is divided in three parts, and shown in Fig 7.

- The main program initialises and establishes a connection with the GPS.
- The other part sets PIN values, PIN directions and calculates readings from sensors.
- The XBee library: the functions defined here enable the Arduino hardware to communicate with the radio.

The same follows for the GS. Software to control GS hardware was written in C++ and runs on a LINUX platform. These executables are in-turn called and run by shell scripts. Fig 8 presents the diagram of the GS software. After start up, the Raspberry Pi opens the UART port and listens for any data from the XBee radio, if there is data, the program reads the data and stores it on a file. The program which monitors system time enters an infinite loop to check the time. If 24 hours have elapsed, the power control script switches ON power to the Iridium modem and the program will commence to transfer data to the Iridium gateway.

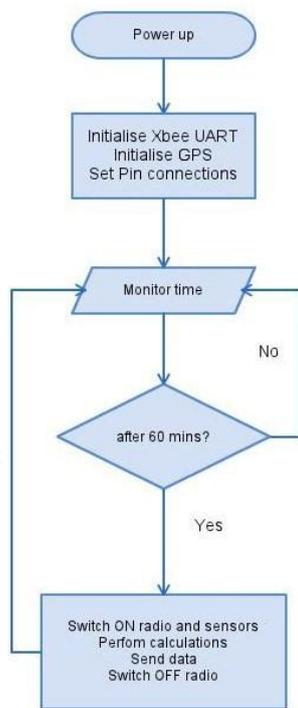


Fig. 7. Sensor node software flow diagram

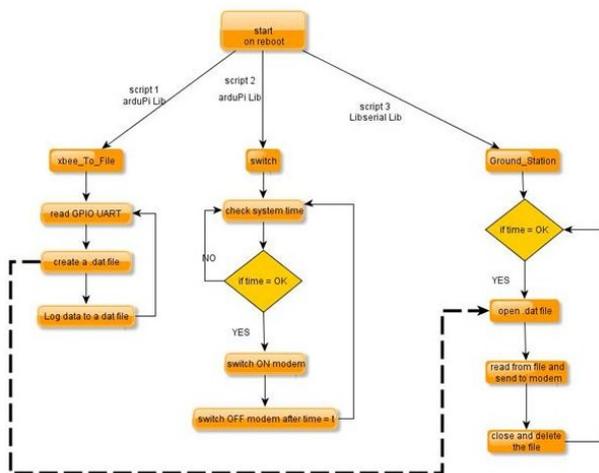


Fig. 8. Ground station software flow diagram

To enable the sensor node(s) to communicate with the sink node, our approach will utilise the star network topology since all the sensor nodes are within the communication range with the sink node. In this topology, all the sensor nodes direct their data to the coordinator which is connected to the GS. In some cases where one or more source nodes are out of range with the coordinator, the nodes use the Ad-hoc on demand Distance Vector (AODV) routing protocol [12] to perform route discoveries to create links thereby extending network coverage and range. If this mode is selected, the network will have to employ accurate time synchronisation algorithms.

VI. RESULTS

We conducted field tests for each of the parameters being monitored. The sensor nodes collect data every 60 minutes and transmit the data to the GS. The GS in turn forwards the data to the MS every 24 hours. Sample results from an actual experiment of wind direction are shown in fig 9 below. The data was sent from the wind vane in a remote location using the developed system.

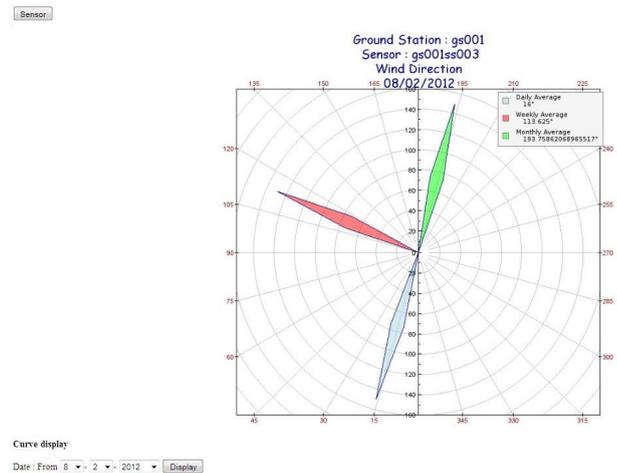


Fig. 9. Polar clock display

Fig 10 below shows the results of an experimental investigation of the power discharge characteristics of the batteries of the sensor nodes and the GS. All the sensor nodes follow an almost similar discharge curve due to the fact that the hardware configuration in the sensor nodes is almost the same. The GS, however, has a unique discharge curve due to a different hardware configuration. The rainfall sensor node achieves roughly 70 hrs of lifetime which satisfies the three day autonomy requirement.

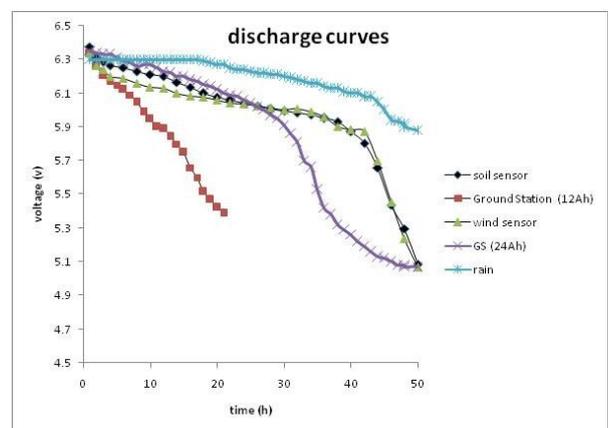


Fig. 10. Battery discharge curves

Fig 11 shows the results of the signal strength of the received packet as a function of the distance between the coordinator and the router. The results from this experiment enable us to characterise the link quality. This can be part

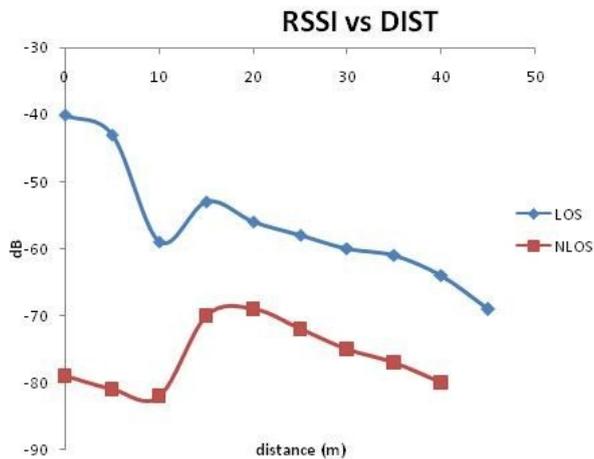


Fig. 11. RSSI vs distance

of future work. The experiment was conducted for Line-of Sight conditions and Non-Line of Sight conditions with only two nodes (one coordinator and one router). The RSSI value for both LOS and NLOS decreases linearly as the distance is increased.

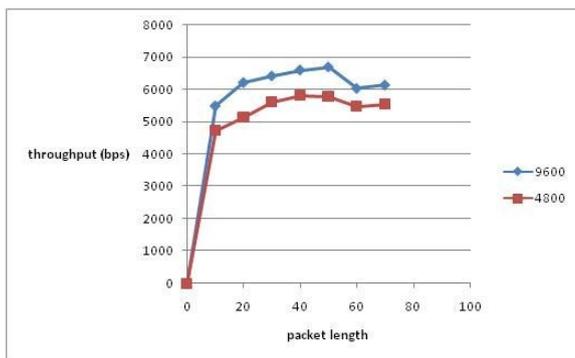


Fig. 12. Throughput for 4800bps and 9600bps

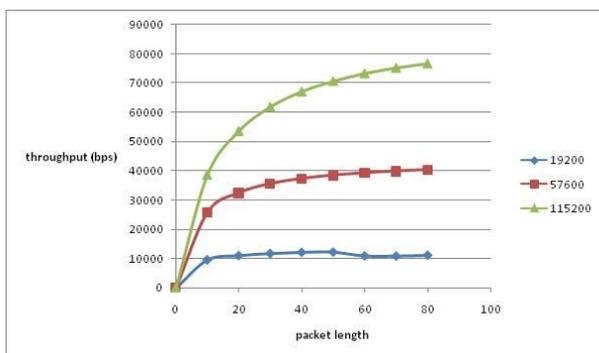


Fig. 13. Throughput for 19200bps, 57600bps and 115200bps

Figs 12 and 13 above present the throughput of the system for 4800bps, 9600bps, 19200bps, 57600bps and 115200bps as a function of packet size. The experiment was conducted with

two nodes directly communicating with each other. The time taken to transmit a packet of n bytes is measured and the throughput is calculated as the ratio between number of bits received and the total transmission time. The experiment is repeated 3 times and the results are displayed in fig 12 and 13. It is noticed that as the speed of transmission is increased, the throughput also increases. We achieved a throughput of 12kbps at a speed of 19200bps and 85kbps was achieved at 115200bps. This could be attributed to a clean channel which was selected by the coordinator and this ensures that data is sent reliably and it is received without the loss of any packets, thereby satisfying the data reliability requirement set out in section III.

VII. CONCLUSION AND FUTURE WORK

In this paper, we have developed a satellite based water resources monitoring system which can be deployed in any remote area and its sole purpose is to provide data in near real time. The project was tested and is functioning well. Overall, the completed system performed satisfactorily as intended and proved its suitability for the particular application. In future, we hope to develop a system which uploads data automatically to the MS. In the current solution, a user downloads data from the email, sorts the data and uploads to the MS.

The power consumption of the system should also be able to benefit from additional development by reducing the solar panel size.

REFERENCES

- [1] Christine Colvin, David Le Maitre, Daleen lotter, 1. "Water Resources and Climate Change Case Study, South African Risk and Vulnerability Atlas", 20:317-330.
- [2] M. A. N2. "Fresh water real-time monitoring system based on wireless sensor network and GSM", *IEEE Conference on Open Systems (ICOS2011)*, pp 354-357, ISBN 978-1-61284-931-7
- [3] S. B. Shamsuddin, M. D. Baba, D. K. Ghodgaonkar, "Design of a Virtual Instrument for Radio Telemetry Station", *Student Conference on Research and development*, pp 414-417, ISBN 0-7803-7565-3.
- [4] Figueredo, A. J. and Wolf, P. S. A. (2009). "Development of an Oceanographic Drifter with Iridium Bi-Directional Communication Capability" *IEEE*, 2012.
- [5] I. F. Akyildiz, E. Cayirci, S. Weilian, "A survey on sensor networks" *IEEE communications magazine*, Vol 40, Issue 8, pp 102-114, ISBN 0163-6804
- [6] Enrique Ramos, "Arduino Basics" *Apress*, pp 1-22, ISBN: 978-1-4302-4168-3
- [7] Atmel Corporation, "Atmega 328P datasheet 2011"
- [8] M. Richardson, S. Wallace, "Getting Started with Raspberry Pi", 2013, ISBN-13: 978-1449344214
- [9] P. T. V Bhuvaneswari, R. Balakumar, V. Vaidehi, "Solar energy harvesting for wireless sensor networks"
- [10] Thiemo Voigt, Hartmut Ritter, Jochen Schiller, "Utilizing solar power in WSN", 2013
- [11] Vongsagon Boonsawat, Jurarat Ekchamanonta, Kulwadee Bumrunghet and Somsak Kittipiyakul, "XBee wireless sensor networks for temperature monitoring", 2005
- [12] A. A Pirzada, M. Portmann, "High Performance AODV Routing Protocol for Hybrid Wireless Mesh Networks", *Annual International Conference on Mobile and Ubiquitous Systems: Networking and Services*, Vol 4, pp 102-114, ISBN 978-1-4244-1024-8, 2007

Comment Classification for an Online News Domain

Dirk Brand
Computer Science Division
Stellenbosch University
7602 Matieland
South Africa
dirkbrand@ml.sun.ac.za

Brink vd Merwe
Computer Science Division
Stellenbosch University
7602 Matieland
South Africa
abvdm@cs.sun.ac.za

ABSTRACT

In online discussion forums, comment moderation systems are often faced with the problem of establishing the value of an unseen online comment. By knowing the value of comments, the system is empowered to establish rank and to enhance the user experience. It is also useful for identifying malicious users that consistently show behaviour that is detrimental to the community.

In this paper, we investigate and evaluate various machine learning techniques for automatic comment scoring. We derive a set of features that aim to capture various comment quality metrics (like relevance, informativeness and spelling) and compare it to content-based features. We investigate the correlation of these features against the community popularity of the comments. Through investigation of supervised learning techniques, we show that content-based features better serves as a predictor of popularity, while quality-based features are better suited for predicting user engagement. We also evaluate how well our classifier based rankings correlate to community preference.

General Terms

Algorithms, Experimentation

Keywords

Regression, Classification, Features

1. INTRODUCTION

There are various online platforms that permit users to generate content. These include forums, blogs, newsgroups and online news providers. The content often has to be moderated for public and corporate benefit. Moderation in the online news domain has recently been a topic of discussion, as users are ever more able to voice their opinions about reported news via some social platform. As the social web grows and people become increasingly socially aware, news sites are becoming ever larger discussion communities where

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...\$15.00.

users can address and comment on common issues spurred by the news articles [1]. One of the key features promoting the success of these online communities is the large-scale user-engagement, seen in the forms of rating, tagging and commenting on content [2]. User-contributed comments offer a much richer source of contextual information than ratings or tags, albeit often a “messy” source of information. Comments are often variable in quality, substance, relevance and style [2].

An online news portal serves many different roles [3]. These rolls fulfil the following tasks:

- educating people,
- providing instant access to the latest news,
- providing feedback for news provider, and
- easily accessible source of information for the general public.

The importance of the roll that online news play in the media sector (especially when educating and informing people) leads news providers to strive to provide content of higher quality. To ensure high quality in user submitted content (such as comments on articles), news providers attempt to moderate or curate the content. Several systems of content moderation have been designed and implemented in the past. These will be explained in Section 7.

2. PROPOSED APPROACH

Previous studies [4, 5, 2] have investigated classification techniques and regression approaches for ranking comments. The authors mentioned above extracted quality-based features from comments using some of the feature extraction techniques mentioned in Section 3. We attempt to show that the same quality-based features are insufficient for predicting a comment’s popularity within the community, but that using only content-based features are better suited (or can at least serve to augment traditional quality-based features). We will also compare the efficiency of the two feature sets for predicting user engagement.

For the quality-based features, we incorporate feature extraction techniques from previous authors. The content-based feature extraction is a new technique for comment ranking (to our knowledge) and seeks to improve on the proposed techniques by instead using a bag-of-words vectors as a feature set. We predict that the quality-based features might be a better predictor of editor preference (over community preference), but the provided data was insufficient to test this hypothesis.

For the supervised learning approaches, a regression filter [6, 7] is applied to a comment and it classifies the comment based on a provided feature set. The regression classifier predicts a continuous numerical value for a comment. We will also investigate the effect of categorising the dependent variable and translating the problem into a classification problem.

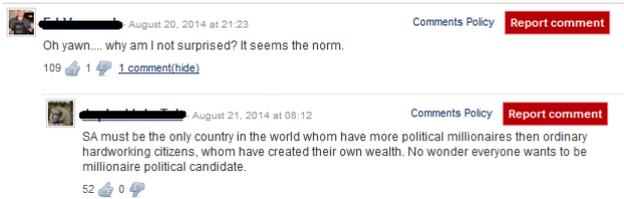


Figure 1: A typical News24 comment thread.

The features will all be extracted from a comment database provided by News24.com (the nature of the data sets are explained in Section 6.1). News24 is a popular South African news provider that allows its users to leave comments on articles. Figure 1 shows a typical comment thread where one user posted a comment and another user commented on his comment (this is an example of the 1-tier commenting that News24 permits). Each article has multiple threads of comments associated with it. These comments will form the basis for our investigation.

3. FEATURE EXTRACTION

Training data is comprised of rows that each consist of a feature vector and an associated value. The model is trained on this training set. The model can then be used to predict the value of a candidate feature vector (a new comment, for instance). The choice of features to use in the training data depends on the domain of the data, as well as the relevance of the features [8].

Consider a set of articles $\{a_1, a_2, \dots, a_k\}$. Denote the i th article by a_i , and its set of n comments by $\{c_{i1}, c_{i2}, \dots, c_{in}\}$. For each comment c_{ij} , a set of features $F_{c_{ij}} = \{f_1, f_2, \dots, f_m\}$ is extracted. The training data then consists of rows of the form $\{(F_{c_{11}}, r_{c_{11}}), \dots, (F_{c_{kn}}, r_{c_{kn}})\}$ where a tuple $(F_{c_{ij}}, r_{c_{ij}})$ indicates a feature set $F_{c_{ij}}$ for comment c_{ij} , and the associated community rating $r_{c_{ij}}$.

We extracted quality-based features, based on previous work. We then explain how a content-based feature set is extracted as a comparison.

3.1 Quality-Based Features

The various features used for the quality-based feature set, are discussed below. The features can be categorised into surface features, lexical features and sentiment features.

• Surface Features

- *Timeliness*. This feature reflects the response time of a user’s comment in relation to when the relevant article was posted [9].
- *Lengthiness*. This feature is a simple measure of the length of a comment relative to the average length of comments of that article [9].

- *Uppercase Frequency*. This feature is a count of the number of words that are completely uppercase [2].
- *Question and Exclamation Frequency*. Both features are the counts of the number of sentences in the comment that end in question and exclamation marks respectively [10]. The values are given as a percentage of the total number of sentences.

• Lexical Features

- *Complexity*. The complexity of a comment is measured by the entropy of the words in the comment [2]. Intuitively, it represents the diversity in word choice in the comment. A low entropy score would indicate that a comment has few or repetitive words.
- *Spelling*. This feature measures the frequency of misspelled words in the comment. The feature is calculated by looking up each word in a dictionary and recording the percentage of words that cannot be found in the dictionary. The dictionary is comprised of words extracted from Peter Norvig’s spell checker data sources [11] and the NLTK [12] sources for male and female names. In future, it would be beneficial to collect data on South African English spelling and use that for the feature.
- *Profanity*. This feature measures the frequency of profane words in the comment. Similar to the spelling feature, the feature value is calculated by looking up each word in a dictionary of profane language and recording the percentage of words that can be found in the list of banned words. The list is built from a list published by Alejandro U. Alvarez [13]¹.
- *Informativeness*. This feature attempts to capture how unique a comment is within its relative thread. The measure that was used, is the standard Term Frequency - Inverse Document Frequency (TF-IDF [14]).
- *Readability*. The readability of a comment is defined as with what ease the reader is able to read the comment (determined by the Flesch Reading Ease Test (FRES) [15]). A high score (above 90) indicates that the text can be understood by an average 11-year old, whereas conversely, a low score (between 0 and 30) indicates that the text will probably only be understood by university graduates.
- *Relevance*. The relevance of a comment can be measured relative to the article or relative to the comment thread that the comment is present in. To calculate the relevance within the comment thread, the overlap between the words in the comment and the words in the article’s comments, is quantified. For this, a bag-of-words vector of the 100 most frequent words is generated from all the

¹The list tries to take common purposeful misspellings of words into account. Eg. ‘butt’ and ‘buttt’ are both in the list. In future, a more domain specific list should be constructed.

comments on an article. Similarly, to calculate a comment's relevance to the article, a bag of words is generated from the body of the article.

• **Social Features**

- *Sentiment.* The text in a comment can be classified as either subjective or objective, and further more as positive or negative (if it was classified as subjective). A trained classifier was used to predict the sentiment of a comment. The classifier was trained and tested with a corpus of 100,000 real tweets (from Twitter ²) that were classified as either positive or negative. The classifier achieved a prediction accuracy of 84.7%.
- *Subjectivity.* The subjectivity of the comment is also captured as a feature. If a comment is between 45% and 55% positive, the comment is classified as objective, otherwise it is classified as subjective.
- *Engagement.* Since News24 uses a one-tier commenting system, users can either leave a new comment ("parent" comment) or comment on an already posted comment ("child" comment). This feature counts the number of child comments to each parent comment.

3.2 Content-Based Features

The above mentioned features attempt to capture the "quality" of a comment. Another way to characterise a comment, is to use the actual content of the comment. To capture this, a list of the most used words in the entire comment space is compiled. Then, for each comment, a vector of the number of occurrences of each word in the comment is created.

For accuracy, stopwords [16] are not consider for the frequent words list. Also, only the stems of words are considered. This is done to group plurals and other word variations into a single representative stem. The Porter stemming algorithm [17] is used (from within the NLTK package [12]) for stemming.

3.3 Value Extraction

The supervised learning methods require a dependent variable (or predictor). Two measures of determining the dependent variable are investigated for this project, engagement and popularity. The engagement that a comment attracts is measured in the percentage of votes on an article, while the popularity is measured by the vote ratio. For the classification methods, the values are discretised into two, three or five balanced categories. The details of the methods are:

- **Percentage of Total Votes** - The ratio of likes to dislikes of a comment: $v = (likes+dislikes)/(\#article\ votes)$.
- **Vote Ratio** - The ratio of likes to dislikes of a comment: $v = likes + c/(likes + dislikes + 2 * c)$. c is a correction term to deal with comments with zero likes or dislikes (set to 5 in our experiments).

²This was chosen as a training set, as it is the closest training set we could find that relates to comments.

4. DATA PREPROCESSING

Various assumptions are made about the training data for the regression and classification models [18]. Firstly, regression (specifically) assumes that each feature is normally distributed (have a zero mean and one unit variance). Secondly, it is assumed that the features are measured without error and are reliable.

4.1 Normalisation

The range of values determined by the above mentioned features varies widely. The regression models that will be considered, all prefer the data to be normalised. If the data is not normalised, it could result in distorted relationships between the features and the value variable [18]. Feature normalisation involves manipulating the feature set to have a zero mean and variance of one.

The goal of standardizing the feature set, is to ensure that the features are in similar ranges. Additionally, standardizing the data allows algorithms such as gradient descent (used in linear regression) to converge faster, and leads to improved performance in algorithms such as Support Vector Regression [19].

4.2 Feature Selection

Reducing the dimensionality of the feature space, results in faster performance for the regression and classification models, as well as a lower variance in the data which means the models can better generalise [20].

A linear regression test is applied to each feature, to establish its F-score [21]. The test works by orthogonalizing the regressor and the data, then computing the correlation between the regressors and finally calculating the F-score. The top K (six in our case) of the features are then selected and are cross multiplied to form $K!$ new features which are then added to the existing data. When the algorithm was run on the quality-based features, it identifies the following six best predicting features: readability, sentiment, subjectivity, thread relevance, timeliness and engagement.

5. REGRESSION MODELS

We apply various regression techniques to determine the predicted community rating of an unrated comment. Regression is a statistical processes to estimate a relationship between variables. In this case, a regression model will help estimate the relationship between a set of features and a score. After a relationship has been estimated, the model can be used to predict the value when presented with a new feature set.

Two regression models will be compared to determine which model best fits the data domain and performs the best, as well as which model gives the highest prediction accuracy. The regression models that will be considered are:

- linear regression [22] and
- support vector regression (with rbf kernel) [23]

The alternative to the regression approach, would be to discretise the continuous value of the regression variable into classes, and using it as input for classification algorithms. For both approaches to determine the dependent variable, the continuous value was binned into sets of two, three and five classes respectively.

As with regression, the classification algorithms are instances of supervised learning techniques that trains on a specified training set of features and classification variables.

Four classification algorithms will be compared and experimented with and evaluated accordingly. The classifiers that will be considered are:

- support vector classification (with rbf kernel) [24],
- support vector classification (with linear kernel),
- logistic regression [25], and
- random forest classification [26].

Both the regression and classification models were implemented with the Scikit-Learn Python library [27].

6. EXPERIMENTS

For all the experiments discussed below, regression, as well as classification with two, three and five balanced classes, are compared. The regression models are scored by doing a 50 fold cross-validation and taking the mean R^2 score [28]³ of all the folds. The classification experiments are evaluated with an accuracy score⁴.

We investigate another measure of predictive accuracy by ranking a list of comments using some ordering. This ranking is then compared to an ideal ordering, determined by ranking the same set of comments by community ratings (like to dislike ratio).

The correlation of the two rankings are measured using normalized discounted cumulative gain (NDCG [29]). NDCG reflects the intuition that accuracy at the top of the list is more important than ranking errors further down the list, which fits the comment ranking model well [2]. NDCG gives a score ranging from 0 to 1, where a higher score indicates a greater correlation between the predicted rank order and the ideal rank order.

6.1 Experimental Setup

For regression, we compare three different types of feature sets and both the vote ratio and percentage of votes are used as the dependent variables. For the classification experiments, only the vote ratio is investigated as a dependent variable.

Firstly, the quality-based feature set, mentioned in Section 3, is used, as well as the features obtained through feature selection (the top six features explain 70% of the variance and are cross multiplied to form extra features). Only comment threads with more than 50 comments are considered. Individual comments are disregarded for the training set, if they have less than five likes or dislikes respectively, less than 50 combined likes and dislikes, or contain less than 100 words. This results in a feature set containing 10296 objects, each consisting of 40 features. The training and test sets make up 67% and 33% of the feature set, respectively.

Secondly, the content-based feature set (a bag-of-words vector) is used. The vector consists of the 100 most used words in the comment space and the frequency at which each comment uses those words.

³a value between zero and one where a higher value represents better predictive accuracy

⁴The percentage of samples correctly classified. Thus, a higher value represents better predictive accuracy

The final feature set consists of the bag-of-words vector, concatenated to the extracted feature vector.

6.2 Results

Feature Set	% of Votes	Vote Ratio
Quality-Based	0.152	0.029
Content-Based	0.032	0.116
Quality + Content	0.150	0.125

Table 1: Linear Regression Result Summary Table.

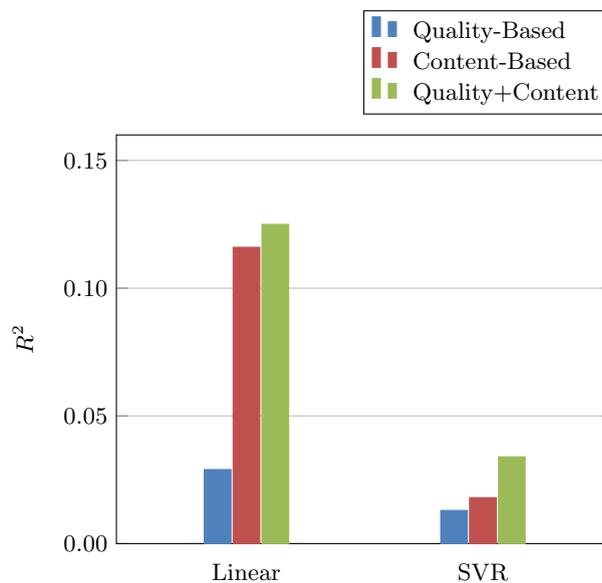


Figure 2: Regression results for training on like-to-dislike ratio.

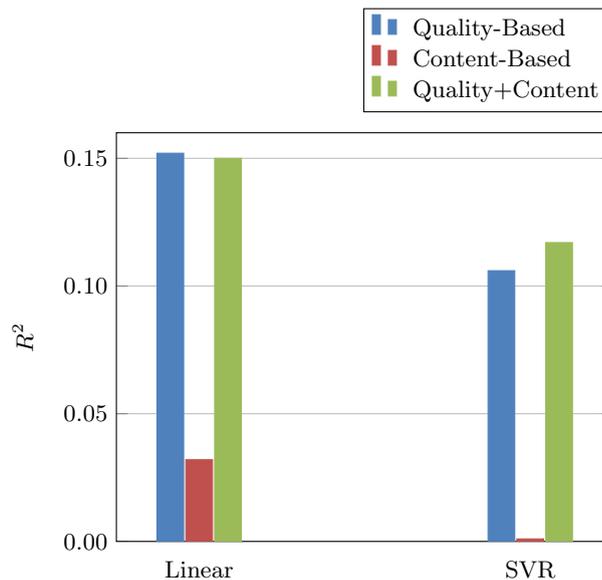


Figure 3: Regression results for training on percentage of total number of votes.

Figure 2 shows that quality-based features are insufficient to predict community preference (when using vote ratio), but that content-based features, as well as quality- and content-based features combined, show better performance.

Figure 3 shows that quality-based features are better suited for predicting engagement (percentage of votes). Augmenting the quality-based features with the content-based features yields similar results.

Algorithm	Two Types	Three Types	Five Types
SVC	0.594	0.425	0.253
Linear SVC	0.616	0.439	0.285
Random Forest	0.588	0.412	0.265
Logistic Regression	0.619	0.438	0.289

Table 2: Results for classification on the quality-based features.

Algorithm	Two Types	Three Types	Five Types
SVC	0.661	0.474	0.314
Linear SVC	0.652	0.466	0.314
Random Forest	0.637	0.430	0.278
Logistic Regression	0.653	0.468	0.309

Table 3: Results for classification on content-based features.

Algorithm	Two Types	Three Types	Five Types
SVC	0.647	0.454	0.313
Linear SVC	0.662	0.455	0.321
Random Forest	0.622	0.410	0.280
Logistic Regression	0.662	0.451	0.323

Table 4: Results for classification on both quality- and content-based features.

The results in Tables 2 to 4 show that the accuracy of the classifier degrades as the data is segregated into more categories, as is expected. It is also evident that, in general, Support Vector Classification performed better than the other models. Table 3 shows that SVC obtains an average accuracy of 47.4% with classifying on the bag-of-words vectors. This is almost as accurate as the classification scores obtained by Wanas et al. [9] (49%), and is deemed sufficiently accurate given the context.

The content-based feature classification clearly outperforms classification on quality-based features, but also when the quality-based features are added to the content-based features.

Training the regression model on total number of votes, rather than the like-to-dislike ratio, results in significantly higher R^2 scores in the regression experiments, indicating that the total number of votes is a better indicator of community preference.

Table 5 shows that the regression accuracy increases logarithmically with the content-based regression model, as the

Vector Size	R^2
50	0.061
100	0.115
200	0.129
500	0.178
1000	0.181

Table 5: Linear Regression on word vectors of different sizes.

size of the vector increases. This shows that even better results are possible with larger vector sizes, but should plateau and diminish when the vectors become too sparse.

6.3 Rank Correlation

Using the trained classifiers, we impose an ordering (or ranking) on a set of comments. This ranking is then compared to an ideal ranking with the NDCG measure.

For the experiment, a linear regression classifier is trained with a training set consisting 19014 comments. The NDCG score is computed with a K value that indicates how many from the list is considered for the comparison (we used $K = 20$). The model is trained and tested with 20-fold cross validation, so the NDCG scores reported in Table 6 is the mean of 20 recorded scores. The classifier is used to predict and rank the list of comments, and the comments' real community like-to-dislike ratio (as in Section 3.3) is used as the ground-truth ordering. NDCG scores range from 0 to 1, where a higher NDCG score indicates that the list ordering in question correlates well to the ideal ordering (i.e. ordered by vote ratio). Table 6 shows that content-based features correlate better to the community ordering.

Quality-Based	Content-Based	Both
0.597	0.782	0.759

Table 6: Normalized Discounted Cumulative Gain with different feature sets for the classifier predicted comments against the community ranked comments.

Further, Figure 4 shows how other orderings compare to our classifier orderings. The 'Random Ordering' simply imposes a shuffle on the comments and runs the NDCG algorithm on the result (intuitively, this should give a lower NDCG score, since the order is arbitrary). The 'Timestamp Ordering' ranks comments in the order that they arrived on the website, with the oldest comment being ranked first (similarly, the order is arbitrary regarding comment popularity, so it should give a lower score).

Figure 4 shows that ordering the comments according to date, or randomly, results in a list that does not correlate well with the community preference, for any feature set. What is encouraging, is that our proposed automatic ranking algorithm performs much better than the other two orderings when the bag-of-words feature vector is used, and according to Table 6, shows comparable performance to the classifier designed by Hsu et al [2].

7. PREVIOUS WORK

Our work in this paper is based on previous studies of comment ranking techniques by Lampe and Resnick [30], Wanas et al. [9], and Hsu et al. [2].

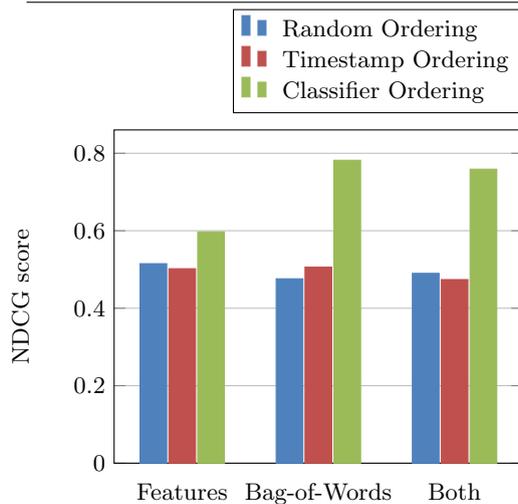


Figure 4: A comparison of NDCG scores for the different feature sets and different list ordering schemes.

7.1 Community Moderation

Lampe and Resnick [30] asked the question: “Can a system of distributed moderation quickly and consistently separate high and low quality comments in an online conversation?”. Their analysis showed that a system that uses the participants in an online conversation as moderators, can efficiently rank comments so as to improve the quality of the conversation. They focused their investigation on slashdot.org.

Firstly, they used the properties of the comments left by users (comment length, word usage), as well as the properties of the authors themselves (frequency of posting, frequency of response) as ways to classify comment. They then found that the judgements of other users were better indicators of which comments needed attention.

Their investigation then involved building a regression model that predicted the final score of an unmoderated comment (what we based our models on), based on the classified comments that the users provided.

7.2 Automatic Scoring and Classification

Wanas et al. [9] sought to improve on the work done by Lampe and Resnick [30]. The latter’s rating system noticed that a significant amount of time had to pass before users could identify good quality comments. Additionally, earlier posts received more attention. Wanas et al. proposed a scheme of automatic post ranking based on supervised learning techniques (Support Vector Classification). Similar work was done by Hsu et al. [2], but using Support Vector Regression.

The features that Wanas et al. used, were based on features designed by Weimer et al. [10], and consisted of various features categorised into five classes. Those classes were relevance, originality, forum-specific, surface (frequency of capitalised words, quality of grammar, etc.) and posting component features (presence and quality of weblinks in posts). The trained classifier designed by Weimer et al. merely classified posts as ‘bad’ or ‘good’ and required that posts used as training data observe proper use of language and linguistics

rules. As observed by Wanas et al, this is not always the case in online forums. They focused their investigation on providing finer ratings for posts, as well as taking various linguistic phenomena that frequent online forums, into account.

Their experiments showed their classifier to be 49.5% accurate when classifying posts as bad, average and good (in terms of their definition of quality). They claim the accuracy to be sufficient to provide rankings for posts. Their experiments also showed that structural features of posts were more significant in classification than features analysing the actual text. This means language independent approaches could be adopted, and led us to investigate improving upon quality-based features with content-based features.

8. CONCLUSION AND FUTURE WORK

The regression and the classification results show that the quality-based features lack in predicting the community popularity of a comment. This could be attributed to biased voting patterns in the community, eg. users that would “like” a comment multiple times if it supports their viewpoint (politically, religiously, or otherwise), but not necessarily evaluate the comment’s quality. Using content-based features performs significantly better and allows us to achieve high comment rank correlation (NDCG) to the community’s preference.

The quality-based features are, however, better suited for predicting the engagement a comment will receive from users in a comment thread.

Future expansions of this research will include designing specific features for the language domain, that incorporate a list of profanities specific to South African English.

The investigated models will also be trained and tested on comments scored by independent editors. We predict that the quality-based features should perform better when predicting editor preference, since it would represent the perceived ordering of comments according to the designers of the commenting system and their desire for what the quality of the comments should be.

9. ACKNOWLEDGEMENTS

Some of this work was done at the MIH Media Lab at Stellenbosch University.

10. REFERENCES

- [1] N. Diakopoulos and M. Naaman, “Towards quality discourse in online news comments,” in *Proceedings of the ACM 2011 conference on Computer supported cooperative work*, pp. 133–142, ACM, 2011.
- [2] C.-F. Hsu, E. Khabiri, and J. Caverlee, “Ranking comments on the social web,” in *Computational Science and Engineering, 2009. CSE’09*, vol. 4, pp. 90–97, IEEE, 2009.
- [3] S. Keibler, “Importance of the online news portal.” <http://www.buddy4study.com/blog/importance-online-news-portal>. Accessed: March 2014.
- [4] M. P. O’Mahony and B. Smyth, “A classification-based review recommender,” *Knowledge-Based Systems*, vol. 23, no. 4, pp. 323–329, 2010.

- [5] M. Rowe, S. Angeletou, and H. Alani, "Predicting discussions on the social semantic web," in *The Semantic Web: Research and Applications*, pp. 405–420, Springer, 2011.
- [6] D. M. Lane, *Online Statistics Education: A Multimedia Course of Study*, ch. Introduction to Linear Regression. Rice University and National Science Foundation, 2004.
- [7] P. University, *WordNet 3.0 Free Dictionary*. March 2014. Statistical Regression.
- [8] M. Karagiannopoulos, D. Anyfantis, S. Kotsiantis, and P. Pintelas, "Feature selection for regression problems," *Proceedings of HERCMA 2007*, 2007.
- [9] N. Wanas, M. El-Saban, H. Ashour, and W. Ammar, "Automatic scoring of online discussion posts," in *Proceedings of the 2nd ACM Workshop on Information Credibility on the Web*, pp. 19–26, ACM, 2008.
- [10] M. Weimer, I. Gurevych, and M. Mühlhäuser, "Automatically assessing the post quality in online discussions on software," in *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pp. 125–128, Association for Computational Linguistics, 2007.
- [11] P. Norvig, "How to write a spelling corrector." <http://norvig.com/spell-correct.html>. Accessed: March 2014.
- [12] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python*. O'Reilly Media, Inc., 2009.
- [13] A. U. Alvarez, "Bad words list." <http://urbanoalvarez.es/blog/2008/04/04/bad-words-list/>. Accessed: March 2014.
- [14] A. Aizawa, "An information-theoretic perspective of tf-idf measures," *Information Processing & Management*, vol. 39, no. 1, pp. 45–65, 2003.
- [15] R. Flesch, "A new readability yardstick," *Journal of applied psychology*, vol. 32, no. 3, p. 221, 1948.
- [16] Ranks.nl, "Default english stopwords list." <http://www.ranks.nl/stopwords>. Accessed: July 2014.
- [17] M. F. Porter, "An algorithm for suffix stripping," *Program: electronic library and information systems*, vol. 14, no. 3, pp. 130–137, 1980.
- [18] J. Osborne and E. Waters, "Four assumptions of multiple regression that researchers should always test," *Practical Assessment, Research & Evaluation*, vol. 8, no. 2, pp. 1–9, 2002.
- [19] R. Herbrich and T. Graepel, "A pac-bayesian margin bound for linear classifiers," *Information Theory, IEEE Transactions on*, vol. 48, no. 12, pp. 3140–3150, 2002.
- [20] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.
- [21] S. MacKenzie, "How to report an F-statistic." <http://www.yorku.ca/mack/RN-HowToReportAnFStatistic.html>. Accessed: April 2014.
- [22] B. Flury and H. Riedwyl, "Multiple linear regression," in *Multivariate Statistics*, pp. 54–74, Springer, 1988.
- [23] D. Basak, S. Pal, and D. C. Patranabis, "Support vector regression," *Neural Information Processing-Letters and Reviews*, vol. 11, no. 10, pp. 203–224, 2007.
- [24] T. Joachims, *Text categorization with support vector machines: Learning with many relevant features*. Springer, 1998.
- [25] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Introduction to the logistic regression model*. Wiley Online Library, 2000.
- [26] V. Svetnik, A. Liaw, C. Tong, J. C. Culberson, R. P. Sheridan, and B. P. Feuston, "Random forest: a classification and regression tool for compound classification and qsar modeling," *Journal of chemical information and computer sciences*, vol. 43, no. 6, pp. 1947–1958, 2003.
- [27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., "Scikit-learn: Machine learning in python," *The Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [28] A. Colin Cameron and F. A. Windmeijer, "An r-squared measure of goodness of fit for some common nonlinear regression models," *Journal of Econometrics*, vol. 77, no. 2, pp. 329–342, 1997.
- [29] W. B. Croft, D. Metzler, and T. Strohman, *Search engines: Information retrieval in practice*. Addison-Wesley Reading, 2010.
- [30] C. Lampe and P. Resnick, "Slash (dot) and burn: distributed moderation in a large online conversation space," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 543–550, ACM, 2004.

A comparison of Quad-tree and Voronoi-based spatial partitioning for dynamic load balancing

M van Greunen
MIH Media Lab
Stellenbosch University
Stellenbosch, South Africa
Email: manrich@ml.sun.ac.za

H.A. Engelbrecht
MIH Media Lab
Stellenbosch University
Stellenbosch, South Africa
Email: hebrecht@ml.sun.ac.za

Abstract—Massively multi-user virtual environments (MMVEs) face scalability challenges, one being the large number of concurrent users interacting in the virtual environment (VE). Spatial partitioning addresses this problem by distributing partitions of the VE, and their associated users, to separate servers. Users dynamically migrate between partitions as they move within the VE and server load imbalances occur when users flock to popular locations (such as cities or boss arenas). Dynamic Load Balancing can be achieved by dynamically scaling the VE partitions and migrating users to underloaded servers. In this paper, we assume an MMVE has load balancing and focus on comparing two spatial partitioning methods, namely Quad-trees and Voronoi diagrams, using OverSim, an extension of the OMNeT++ simulation package. We evaluate each approach using the number of messages sent between servers, the distribution of users across servers and the number of servers in use as performance metrics. We conclude that a Voronoi based system is better in distributing the load across multiple servers, but has a greater computational cost than a Quad-tree based system.

I. INTRODUCTION

Interest in Massively multi-user virtual environments (MMVEs) has increased dramatically over the past few years, particularly in multimedia and online gaming. MMVEs enables multiple users to assume virtual representations called *avatars* to interact and socialise inside a virtual world. Client/Server architectures are predominantly used for providing the resources required to host an MMVE from a central location. However, as the number of concurrent users of MMVEs increase, a single server is not capable of reliably providing the MMVE functionality. One solution to address this problem, is to distribute the load of hosting the MMVE to server clusters. Companies that develop and host MMVEs (such as *World of Warcraft* and *Diablo 3*) invest in large server infrastructures to provision for dynamic loads. Cloud Computing (or On-Demand Computing) now enables independent game developers, especially South African game developers to develop and host MMVEs on a distributed server cluster in the cloud, without the need for large upfront investments. The cloud could dynamically scale according to the number of users, ie. the load of the MMVE and be able to operate at the scale of other AAA game development companies' titles.

Using distributed server clusters introduces a number of challenges, such as how to distribute the virtual environment (VE) and keep the state of the MMVE consistent amongst the server nodes. Prominent approaches that address these

challenges include *sharding*, *zoning* and *spatial partitioning* of the virtual environment. Sharding of the VE entails that the VE be duplicated and hosted on individual servers, whereby zoning entails that separate regions of the VE be hosted on separate servers and these regions be connected with so called 'portals' inside the virtual world. Both of these methods has a disjointed VE, with inter-server latencies causing users, connected to different server nodes, unable to interact and modify the game state in a *consistent* way.

Spatial partitioning divides the virtual environment (VE) into separate regions. Each region is hosted by a single server and that server manages the state of all entities within its region. Entities can move within the VE and are assigned to a new server node when crossing the border between regions, effectively migrating the state of the entity. An avatar can directly modify the state of an entity or indirectly cause the state of an entity to be modified. Both direct and indirect interactions consume resources on the server node, in terms of bandwidth and computational power. In this paper we do not make a distinction between direct and indirect interactions, but as a simplification, regard the number of avatars in a region as indicative of the load of that server node. By varying the size of regions assigned to each server node, the total load of the server cluster can be distributed. We investigate two spatial partitioning algorithms, namely Quad-trees [1, p. 307] and Voronoi diagrams [1, p. 147], for partitioning the VE in a Dynamic Load Balancing (DLB) system. The particular contributions of this paper is in the evaluation and comparison of the load balancing efficiency of both algorithms in terms of network bandwidth, server utilisation, user count and computational cost.

II. RELATED WORK

Yu and Vuong [2], developed a Mobile Peer-to-Peer Overlay Architecture (MOPAR) for interest management of MMOGs that use *zoning*. By dividing the virtual environment into same-sized hexagonal zones and assigning each to a server node, the virtual world is distributed and scalability is achieved. By utilising an Area of Interest (AoI) neighbouring scheme together with virtual locations, each node only handles updates from its neighbours. They focused on guaranteeing all clients have consistency across zones within the virtual environment [2]. Yu and Vuong do not show experimental or simulated results, but their analysis only states the following: "MOPAR uses less nodes, uses resources more effectively

and is more fault-tolerant than other Peer-to-Peer interest management schemes.” Backhaus and Krause [3], proposes QuON: a Quad-tree based Overlay Protocol for distributed virtual worlds. QuON sends game event messages to a client’s neighbours, defined by an AoI. Because update-messages are only sent to nodes which will be affected, the load is distributed among all peers. State consistency and security is achieved by organising neighbours into Quad-trees. QuON is shown to be a practical and effective solution over C/S networks based on OverSim simulation results [3]. Hu and Chen [4] presents VSO: a Voronoi-based Self-organising Overlay (VSO), which dynamic balances the load using spatial partitioning. Neighbouring nodes are organised into Voronoi diagrams and using Spatial Publish/Subscribe (SPS) clients are able to subscribe to a Voronoi region that is the responsibility of a server node. VSO measures the load of a specific node, based on the number of subscribers to that node’s region. When a specific node is overloaded it requests a neighbouring node to move its virtual location to resize the enclosed area. Clients are unsubscribed from the overloaded node and ownership is transferred to neighbouring nodes. VSO evaluations shows an improvement in bandwidth usage of up to 95% when comparing static and dynamic partitioning for up to 500 nodes. In terms of scalability VSO shows that for 1000 nodes, 90% of servers are still functioning under an ‘acceptable load’ as specified by Hu and Chen [4].

III. SYSTEM MODEL

The system model of an MMVE consists of multiple users that dynamically connect to a VE and generate events that include the movement of their avatar, the manipulation of VE objects and interaction with server-controlled entities (called non-player entities or NPEs) [5]. These events are sent to the server which updates the global game state using server-side game logic, generating state updates, that are then distributed to all subscribed clients. According to Carlini et al. [6], the management and handling of events created by user avatars and passive entities, constitutes the typical computational and bandwidth load of a distributed VE.

As previously mentioned, spatial partitioning divides the VE into regions and distributes these regions to server nodes that are responsible for the game state of that region. A client/server architecture that supports spatial partitioning is shown in Fig. 1. It consists of four parts: (1) distributed servers that are connected in a server cluster, (2) a partitioned VE (3) clients and (4) a directory server.

The **Client** is a representation of a user within the VE. It has a virtual location within a particular region and the server hosting that region is responsible for handling its state updates. A client can change its virtual location incrementally and is *migrated* when it moves across a border between regions. *Migration* is defined as the transfer of the responsibility of a client from one server to another. Migration requires communication between the servers and the migrating client so that state can be transferred and the state of the VE kept consistent.

A **Server** is responsible for hosting a region within the VE. It has a virtual location, a list of clients it owns and a list of server neighbours within the VE. A server can measure its

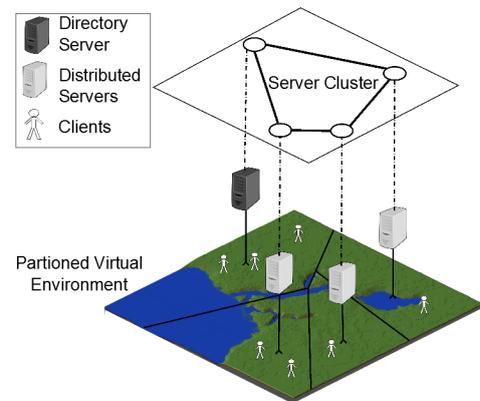


Fig. 1. The spatial partitioning client/server architecture: Distributed servers are connected in a Server Cluster and each is responsible for a region in the Partitioned Virtual Environment and all Clients in that region.

load and determine its *ownership* of a specific client, defined as that client being within the server’s region.

The **Directory** server is used as the entry point for clients connecting to the VE and for new servers joining the distributed server-cluster. Initially the *Directory* server is the only server and it owns the entire VE. The typical role of a directory server is to verify users accounts and redirect clients to servers hosting the VE, but for the purpose of this paper we assume that this does not have a significant contribution to the load.

IV. DYNAMIC LOAD BALANCING

Dynamic Load Balancing require that the regions can be dynamically resized and that clients can be migrated to underutilised servers. We define two states of server load, namely **overloaded** and **under-loaded**. For the purpose of this paper, a server is deemed to be overloaded when the number of clients it owns exceeds a predetermined overload threshold. A server is deemed to be under-loaded when the number of clients it owns is lower than the under-load threshold.

When an active server enters the overloaded state, it migrates its load by setting up a new server. The new server is provided with a virtual location and region to host. All neighbouring servers are also notified of the new server so they the servers an adjust their regions to accommodate the region of the new server. Clients that are now within the region of the new server is migrated to that server. When an active server enters the under-loaded state it returns the region it is hosting to its neighbouring servers. Clients are migrated to the neighbouring server hosting the region where its avatar resides.

There are two issues that still need to be addressed for dynamic load balancing: (1) distributed computation of the VE partitioning and (2) implementation of dynamically load balancing by resizing and migrating VE regions between servers. Both of theses issues will be discussed when we discuss the detail of Quad-trees and Voronoi diagrams in the rest of this section.

A. Distributed computation of Quad-trees partitioning

A Quad-tree is a rooted data structure starting with the first (parent) node and is characterised by each parent node

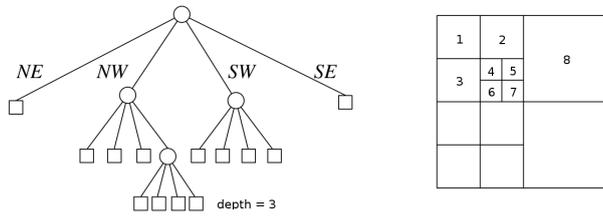


Fig. 2. The Quad-tree structure [7].

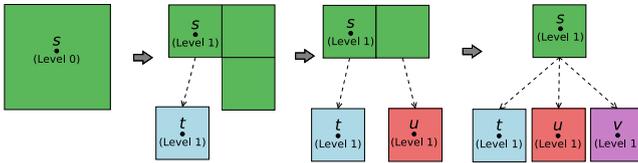


Fig. 3. Quad-tree partitioning of the VE. Partitioning level limited to level 2.

having four child nodes. In the case of two-dimensional VE partitioning, each server node owns a square in the VE, with each child node of that server responsible for a quadrant of its square (see Fig. 2). Using the recursive property of Quad-trees one can construct an algorithm for computing a Quad-tree as follow. For any server node n , divide its square into four squares, calculate each square with side length half of n 's original side length and find all neighbours of n .

This same property can be used to distributively subdivide each server node's own square. In order for the global Quad-tree structure to be kept intact, a *Parent-child relation* is defined by: each distributed child node keeping a reference to its parent node (i.e. the server node that created it).

Quad-tree neighbours are typically defined as server nodes that share an edge [1, p. 313], but in this paper we include neighbours that share a vertex. In order for the child node to determine its neighbours, it evaluates its parent's neighbour and child nodes. Fig. 2 illustrate node 4's edge neighbours as, $\{2, 3, 5, 6\}$. But including node 4's vertex neighbours it becomes, $\{1, 2, 3, 5, 6, 7\}$. Server node 1 should include server node 4 as a neighbour, because the squares regions represent the VE and if a user's avatar move diagonally (from 1 to 4) across the border, it should be migrated.

B. Quad-tree DLB implementation

To implement dynamic load balancing servers need to be able to migrate VE regions, which is achieved by using the previously mentioned *Parent-child relation* as shown in Fig. 3. Each *Child* server keeps a reference to the *Parent* server that created it. The *Directory* server s is the root node and first *Parent* node of the Quad-tree and is initialised with the entire VE at level 0. When the *Directory* server s determines that it is overloaded, it partitions its region into four squares and sets up a new child server t , increasing its own level and assigns one of its four squares to t . If s becomes overloaded again, it sets up a new child server u and assigns it one of its squares regions. This procedure continues until s owns just one smaller square region. This Quad-tree partitioning scheme is one of many possible methods. Some of these methods allows for

regions to be made-up of a combination of squares (including different level squares). This creates complex regions and intricate neighbouring schemes. Our approach is similar to the one used in Backhaus's QuON [3] and makes for a simpler neighbouring scheme.

In our Quad-tree DLB system we limit the partitioning level to 2, preventing the number of neighbours to increase indefinitely, but be limited to a maximum of eight neighbours. The Quad-tree in itself has an unlimited number of levels. A *Server* discovers its neighbours by determining if it shares an edge or a vertex with any of its parent's neighbours or other child servers.

C. Distributed computation of Voronoi diagrams

A Voronoi diagram in two-dimensions, is a tessellation of a plane into a set of polygons V_1, \dots, V_n , associated with the nodes v_1, v_2, \dots, v_n respectively. These polygons are termed Voronoi cells [8] and are calculated using the following (taken from [1, p. 149]):

1. For two nodes v_p and v_q and their associated locations p and q in the plane we calculate the *bisector* of p and q as the perpendicular bisector of the line segment \overline{pq} . This gives two half-planes, denoted by $h(p,q)$ and $h(q,p)$ containing the points p and q respectively.
2. Calculate $n-1$ half-planes where n is the number of nodes.
3. The half-plane intersections define a region that is bounded by a convex polygon, with at most $n - 1$ vertices and at most $n - 1$ edges.

Algorithms for calculating Voronoi diagrams that use the coordinate information about *all* nodes in the plane and are termed full information algorithms. The Voronoi diagram can be calculated in a distributed manner by each server node v_i using a full information algorithm such as Steve Fortune's sweep-line algorithm [9], but limiting the calculation to only using the server's own VE location and the VE locations of its *relevant* neighbours. A *relevant* neighbour for server node v_i is defined as a node that will change v_i 's Voronoi cell if included in server v_i 's Voronoi diagram calculation.

D. Voronoi DLB implementation

The Voronoi DLB implementation is best explained using a typical case as shown in Fig. 4. If s is overloaded, it sets up a new server t and assigns it a location within the VE. For the purpose of this paper, we chose the virtual location as the geometric centroid of the locations of s 's clients. Server s should transfer some of its clients to t and reduce its own load, thus we assume that choosing the client-locations centroid will suffice in satisfying this requirement. If s becomes overloaded again, a new server u is set up and both s and t determines if u is a *relevant* neighbour and recalculates their Voronoi cells. To determine the *relevant* neighbours we first need to determine the set of possible neighbours from which the *relevant* neighbours is a subset. Fig. 5 illustrates our neighbouring scheme: Server s owns a Voronoi cell $V(s)$, calculated using s 's location and its current neighbour locations $\{t, u, v, w\}$. We define a circle centred at s with radius r , where r is the maximum distance to any vertices p in $V(s)$.

$$r = \max_{p \in V(s)} [dist(p, s_{loc})] \quad (1)$$

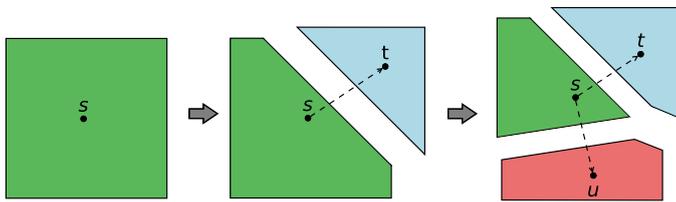
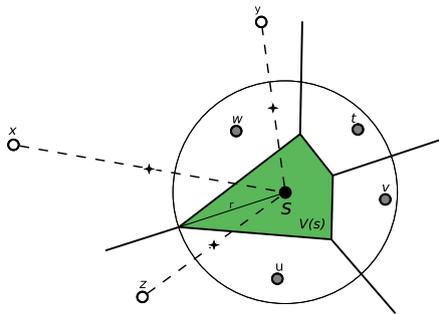


Fig. 4. Voronoi partitioning of the VE.


 Fig. 5. Voronoi neighbouring scheme: Determining if x , y , or z are possible neighbours of s .

Next, s tries to eliminate x , y , or z as possible neighbours, by calculating the midpoint between it and the possible neighbour. If the midpoint is outside its circle with radius r , the server is discarded as a possible neighbour, e.g. in the case of x . If the midpoint is inside s 's circle, the server is kept as a possible neighbour.

The same test is executed for y and z and both results in s including them as possible neighbours. When referring back to the definition of a *relevant* neighbour, a server node that will affect $V(s)$, y should be excluded from s 's list of neighbour. Up to this point, s has no way to determine whether both y and z are *relevant* neighbours, but as Cao [8] stated, including y as an additional node does not affect the Voronoi calculation of $V(s)$ except by increasing computational costs. For the purpose of this paper we will accept the additional cost of including irrelevant neighbours.

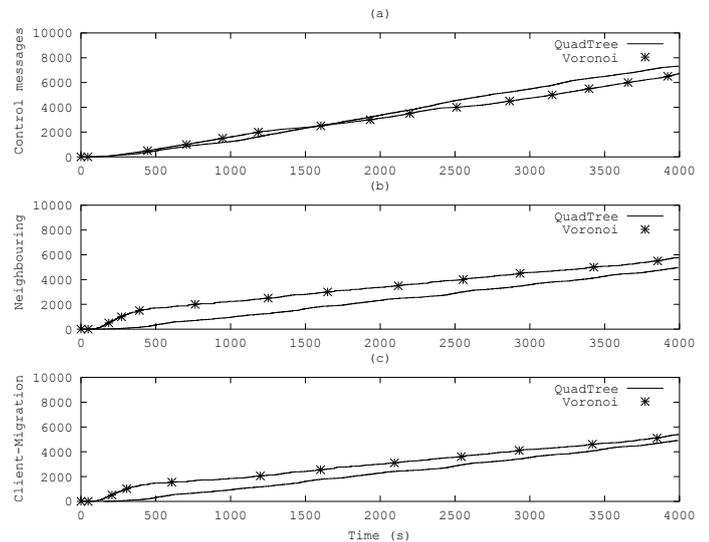
V. EVALUATION

In this section we present the evaluation of both DLB systems. The purpose of this evaluations is to determine which of the DLB systems is more efficient at balancing the load across a distributed VE. We use the following to evaluate the performance of the DLB systems i.e. *bandwidth*, *effectiveness*, *fairness* and *partitioning computational cost*.

A. Metrics

Bandwidth: Network bandwidth is measured as the number of inter-server and server-client messages sent and the scheme with less messages requires less bandwidth and thus increases the MMVE user experience. Inter-server messages are divided into *Control* and *Neighbouring* messages while *Client-Migration* messages are server-client messages, i.e.:

- Control messages are used to indicate a overloaded server, under-loaded server or to request a new server be started.
- Neighbouring messages are used to notify about neighbour


 Fig. 6. Quad-tree vs. Voronoi: The number of (a) *Control*, (b) *Neighbouring* and (c) *Client-Migration* messages.

changes.

- *Client-Migration* messages are used to migrate a client to a neighbouring server.

Effectiveness: The effectiveness of the DLB scheme is determined by the distribution of the number of clients owned by each server.

Fairness: The number of servers in use, overloaded or available (i.e. under-loaded) indicates the *fairness* of each DLB system. A *fair* system is one where the servers in use have a similar load.

Computational cost: The per server computational cost of each partitioning algorithm gives an indication of the additional computational burden of the DLB scheme.

B. Simulation parameters

Simulations are performed in OMNeT++ with OverSim [10], a discrete event simulator for peer-to-peer networks which simulates real-world network conditions and enables simulations to be executed on physical server clusters. We set the VE dimension to $10,000 \times 10,000$ units and a client's movement speed is set to 10 units/second, which is comparable with typical MMVE environments and movement (i.e. *World of Warcraft*). Clients are added to the VE at the *Directory* server's location, every 5 seconds up to a maximum of 40 clients. We define overload threshold as 5 clients or more. The number of available servers are limited to 10 instances. The length of simulation is set to 4,000 seconds. We assume this to be a sufficient time for the system to become fully loaded and reach a steady state. We use Pastry as a peer-to-peer communication overlay to simplify network communication, but for the purpose of the simulation the specific communication scheme is not important.

C. Simulation Results

The network bandwidth results for both DLB schemes are shown in Fig. 6. Fig. 6(a) shows that both DLB schemes send a similar number of *Control* at the beginning of the simulation.

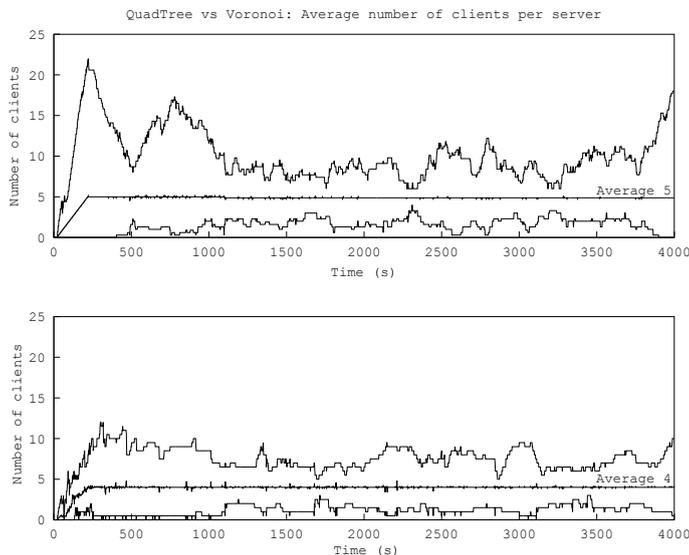


Fig. 7. Quad-tree(top) vs. Voronoi(bottom): The average number of clients per server.

At 2500 seconds the Quad-tree scheme's *Control* messages increase, possibly indicating that the Quad-tree servers are overloaded. The reason it's more than the Voronoi server's, could be because of the limitations on our Quad-tree structure, only being divisible to level 2 and thus causing a server to remain overloaded when reaching this level. Fig. 6(b) shows significantly more *Neighbouring* messages for Voronoi DLB up to 400 seconds as compared to Quad-tree DLB. This can be contributed to Voronoi's dynamic behaviour to resize regions in response to the changing load. The increase in *Client-Migration* messages seen in Fig. 6(c) can also be attributed to this rapid response to load increase. It can be seen that Voronoi DLB results in less *Control* messages being sent, but that more *Neighbouring* and *Client-Migration* messages are needed that for Quad-tree DLB.

Fig. 7 shows the minimum, maximum and the average number of clients per server over time, for both Quad-tree (top) and Voronoi (bottom) DLB systems. We observe that Quad-tree DLB has a higher maximum users count per server and a larger spread compared to the Voronoi DLB system. The average number of five clients per Quad-tree server is also higher than the average of four clients per server for Voronoi DLB. This indicates that Voronoi DLB is more effective than Quad-tree DLB, since all Quad-tree servers are not being utilised. However, this could be as a result of the previously mentioned limitation imposed on the Quad-trees. We also note that Voronoi DLB is more fair than Quad-tree DLB, since the spread of the number of clients per server is smaller.

Table I shows the *Computational cost* of each partitioning scheme. We have counted the number of computations, measure the calculate-time of a single partitioning and the total time spent calculating partitioning during the simulation. To ensure accuracy of the results, each algorithm is run 10,000 times and the average time used. We observe much higher computational cost for calculating Voronoi diagrams as compared to calculating Quad-tree partitioning. The Voronoi partitioning algorithm takes longer and is calculated more frequent than Quad-tree

Algorithm	Comp. time	No. of calc's	Total cost
Voronoi	33 ms	4953	163.45 s
Quad-tree	0.96 ms	29	28 ms

TABLE I. COMPUTATIONAL COST.

partitioning. This results in Voronoi DLB having a computation cost of 5837.5 times more than Quad-tree DLB. We deem the average computational time of 33ms to be acceptable, since Carlini et al. states that the computational cost of a partitioning algorithm should simply be acceptably low [6]. We assume that an acceptable range is less than 50ms, half of the typical acceptable MMVE of 100ms [11].

VI. CONCLUSIONS AND FUTURE WORK

In this paper we have shown that, for dynamic load balancing of distributed MMVEs, using Voronoi diagrams for spatial partitioning is significantly more computationally expensive than Quad-trees, and also use more network bandwidth. However, Voronoi-based DLB system is more effective in utilising the available servers, because it can assign dynamic sized regions in comparison to the Quad-tree DLB system which has limitations on when a region can be divided as well as returned to a parent node. It has also been shown that Voronoi DLB is more fair, since it results in an average of less clients per region.

In future we plan on enabling Voronoi-based servers to move their virtual location to resize the region they own, as Hu and Chen [4] proposed as well as investigate the effect on Quad-tree effectiveness when allowing Quad-tree based servers to subdivide their regions into more levels. We also want to verify the simulation results by implementing both DLB schemes in a real MMVE such as Minecraft.

REFERENCES

- [1] M. de Berg, M. van Kreveld, O. Cheong, and M. Overmars, *Computational Geometry, Algorithms and Applications*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008.
- [2] A. Yu and S. Vuong, "MOPAR: a mobile peer-to-peer overlay architecture for interest management of massively multiplayer online games," (*NOSSDAV'05*), pp. 99–104, 2005.
- [3] H. Backhaus and S. Krause, "QuON: a quad-tree-based overlay protocol for distributed virtual worlds," *Int. Jo. of Advanced Media and Communication*, vol. 4, no. 2, p. 126, 2010.
- [4] S.-Y. Hu and K.-T. Chen, "VSO: Self-Organizing Spatial Publish Subscribe," *2011 IEEE 5th Int. Conf. on Self-Adaptive and Self-Organizing Systems*, pp. 21–30, Oct. 2011.
- [5] J. S. Gilmore, "A state management and persistency architecture for peer-to-peer massively multi-user virtual environments," Doctor of Philosophy, Stellenbosch University, 2013.
- [6] E. Carlini, L. Ricci, and M. Coppola, "Flexible load distribution for hybrid distributed virtual environments," *Future Generation Computer Systems*, vol. 29, no. 6, pp. 1561–1572, Aug. 2013.
- [7] K. Buchin, "Geometric Algorithms Lecture 9 : Quadtrees," 2013.
- [8] M. Cao and C. Hadjicostis, "Distributed algorithms for Voronoi diagrams and application in ad-hoc networks," *Preprint, Oct*, pp. 1–12, 2002.
- [9] S. Fortune, "A sweepline algorithm for Voronoi diagrams," *Algorithmica*, vol. 2, no. 1-4, pp. 153–174, 1987.
- [10] I. Baumgart, B. Heep, and S. Krause, "OverSim: A flexible overlay network simulation framework," in *Proc. of 10th IEEE Global Internet Symposium (GI '07) in conjunction with IEEE INFOCOM 2007, Anchorage, AK, USA, May 2007*, pp. 79–84.
- [11] A. Yahyavi and B. Kemme, "Peer-to-Peer Architectures for Massively Multiplayer Online Games : A Survey," *ACM Computing Surveys*, vol. 46, no. 1, 2013.

A Quantitative Measure of Congestion in Stellenbosch using Probe Data

Dominique ter Huurne and Johann Andersen, Stellenbosch Smart Mobility Lab (SSML), Department of Civil Engineering, *University of Stellenbosch, South Africa**

Abstract—This paper aims to quantify and evaluate congestion in Stellenbosch, a historic university town located approximately 50 kilometres east of Cape Town, South Africa, using probe data. It is known that Stellenbosch experiences traffic congestion, but the scientific extent of this congestion has not been fully determined, as the present volume counts alone are not a sufficient form of assessment. Its residents complain about congestion suffered in town and express frustration. This, along with the fourth annual TomTom South African Traffic Index publication, which revealed that Cape Town (with a congestion index of 27%) is the most congested city in South Africa, instigated this study. Literature bares that the level of service concept (LOS) defined in the Highway Capacity Manual (HCM) has been widely used as a basis for congestion measures, although travel-time-based measures are suggested to satisfy the need for congestion information best. Travel time is well understood by both the general public and professional community, but the collection of travel time, travel speed, travel rate and travel delay data is historically deemed somewhat more complex and onerous than traffic volume counting procedures, and together with limited financial resources has restrained its application. The methodology applied in this study comprises the utilisation of TomTom Traffic Stats Portal that contains historic travel-time-based data from TomTom in-vehicle navigation systems and supporting devices. The platform and associated configuration is state-of-the-art and brings new light to travel-time-based congestion measures. The data was statistically analysed over various date and time periods, and standard congestion index concepts were applied. Congestion measures were considered along the major arterials leading into and out of Stellenbosch, as well as on part of its central road network. This paper shows that Stellenbosch evidently faces increased levels of congestion. Travel times on the inbound arterials are on the rise, and in-town traffic is becoming unsustainable.

Keywords—TomTom; probe data; congestion measurement

I. BACKGROUND

According to the 2011 census, Stellenbosch Municipality, Western Cape, South Africa, (governing the towns of Stellenbosch, Franschhoek, Pniel and surrounding rural areas) has a population of 155733, and 43200 households covering an area of 831km² [3]. The town Stellenbosch has a surfaced road network of 235777m, and 0.9 private cars per household, according to a household survey conducted in 2008 [5]. Stellenbosch is home to University of Stellenbosch with approximately 28156 enrolled students and a personnel size of around 3085 (2013) [4]. Over the ten-year span from 2004 to 2013, the number of students increased by 28%. A study conducted by the Stellenbosch Municipality in 2009 reported that one third of the students reside in or near campus; another third reside in the town or the immediate surrounding area; and the final third reside in the surrounding towns or the Cape Metro

* Miss ter Huurne is a full-time Masters (research) student at the Department of Civil Engineering at the University of Stellenbosch, conducting her research in ITS (email: 15782492@sun.ac.za).

Dr. S. J. Andersen is an Industry Associate Professor in Intelligent Transportation Systems at Stellenbosch University (email: jandersen@sun.ac.za).

[5]. Furthermore, 51% of the students use the passenger vehicle as their mode of transport to and from campus, of which 85% are also the driver of the vehicle [5]. Of all the personnel, 83% use the passenger vehicle to work daily and 87% of them are also the driver [5]. 27 schools are located in Stellenbosch, spread across the various suburbs and township. 8 of these schools are high schools, attracting learners from neighbouring towns and even other parts of the country.

II. INTRODUCTION

Congestion (and its associated bottlenecks) is observed in Stellenbosch on a daily basis, and results in complaints and frustration expressed by its residents. The fourth annual TomTom South African Traffic Index publication revealed that Cape Town is the most congested city in South Africa of late, with a congestion index of 27%. The Stellenbosch Smart Mobility Lab (SSML) deemed that this necessitated a quantitative measure of the true extent of the congestion in Stellenbosch, located only 50km east of Cape Town, which goes beyond volume counts and personal perception. Congestion is here defined as a condition that occurs on roadways as the demand increases to its carrying capacity, and the number of vehicles arriving is greater than the number of vehicles discharged. It is characterised by slower speeds, longer travel times and increased vehicular queuing. Two methodologies, namely the Level of Service Concept and Use of Probe Data, are explained and compared before the use of TomTom probe data is carried through in the rest of the paper. The compilation of the TomTom datasets/queries is discussed, after which congestion indices (speed reduction index and congestion index) and other congestion measures (travel rate, delay rate, relative delay rate and delay ratio) are explained and applied to the given output. An evaluation of the resulting numerical values finally follows.

III. METHODOLOGY

A. Methodologies used in congestion measurement

1) Method of the past: level of service concept

The Highway Capacity Manual uses the level of service (LOS) concept to represent a range of roadway operating conditions. This concept has been widely applied to congestion measurement [2]. A shortcoming of the LOS technique is its use of letter grades in place of a numerical scale, and that there is no consensus regarding the LOS range corresponding to the threshold of congestion [2]. It also gives no detailed subclassification within LOS F (worst condition). Although most congestion management agencies commonly used the LOS concept as their measure of congestion, delay and travel time/speed were the suggested measures for use by most agencies [2]. The most frequently cited reason for not using delay and travel time/speed was limited financial resources, as

data collection techniques such as the floating car and licence plate matching were used before probe data became available [2].

2) *State-of-the-art method: use of probe data*

Probe data is information amassed while monitoring a sample of transportation system users as they pass predefined points along a segment of thoroughfare. TomTom probe data sources include connected GPS devices, GSM devices, road sensors and incident data. In this paper, only motorised transportation is considered, but ideal probes span multiple modes of transit. Probe data has the advantage that it is more accurate and/or less expensive than most current data collection devices and techniques; and as a non-infrastructure solution, it avoids the following predicaments: theft/vandalism, collisions, communications, power, etc. Although the field of Intelligent Transportation Systems (ITS) exists for a number of decades already, the use of probe data only intensified recently (last decade). Probe data finally enabled professionals to measure a fundamental performance indicator, travel time, readily and with greater precision.

B. *Methodology of this paper*

1) *Specification of routes*

7 routes were studied for this paper (in both directions). These are shown in Fig.1, with their lengths given in brackets. These routes are the major arterials leading into and out of Stellenbosch, and also some of the interior roads linked to these arterials and observed to be exceptionally congested.

2) *Specification of analysis date and time periods*

Probe data was collected using the TomTom Stats Portal containing historic data. The date period was set to a typical day (Tuesday to Thursday, February to the end of March) for the years 2011 to 2014. 7 time periods were selected for each day of the defined date period. These are: (1) 12am to 6am (free flow), (2) 6am to 7am, (3) 7am to 8am, (4) 8am to 9am, (5) 1pm to 3pm, (6) 4pm to 5pm and (7) 5pm to 6pm.

3) *Analysis of the output*

The TomTom Traffic Stats Portal generates 4 output formats for each submitted dataset. These are (1) a KML file, (2) a XLS file, (3) a shapefile and (4) charts that open in the portal. These outputs provide segment, speed and travel time information. It was ensured that all sample sizes are adequate. Where the sample size is below 50 for the comparative time periods and below 10 for the base period, comparisons were made to the previous year to assess the correctness of the output. Most comparative-time-period sample sizes lie between 100 and 600.

To obtain a general overview of the congestion level in Stellenbosch, the peak-hour delay was computed for each route, for each analysed year from the obtained outputs. Delay is here defined as the difference between the actual travel time and free-flow travel time, and is a simple, easily-understood measure for attaining a first impression. The typical delays are shown in Table I, with two non-typical time periods included for 2013. (The date period was once altered to include Fridays, i.e. Tuesday to Friday, and then modified to the June/July school and university vacation period.) It was immediately

evident that the 6-7am time period provides no pertinent data and was thus ignored.

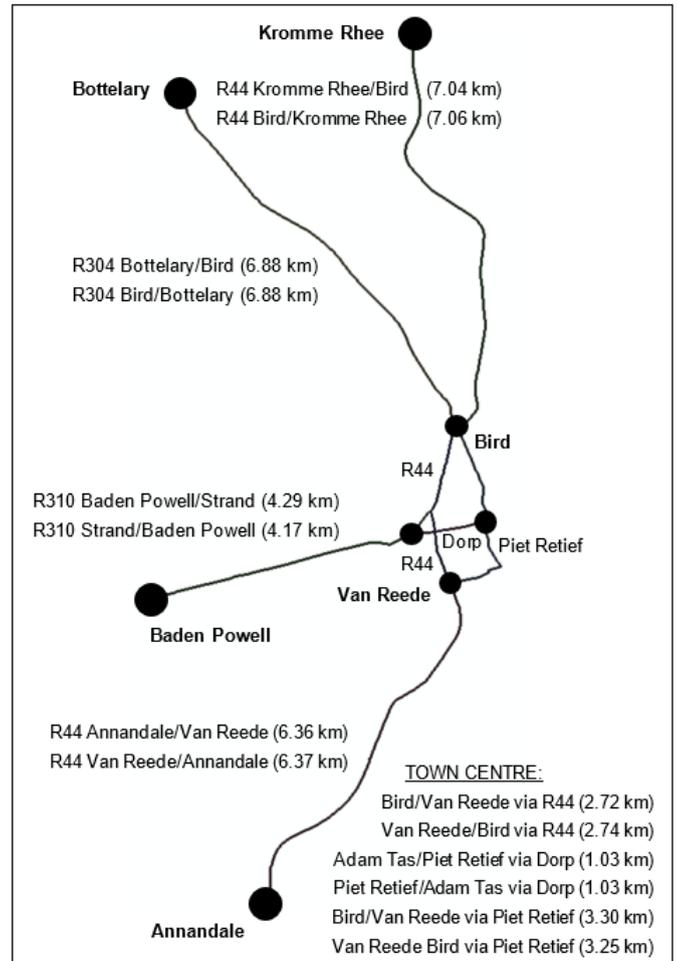


Fig. 1. Map of selected routes.

The more complex and scientific congestion indices and measures are enlightened below.

a) *Speed reduction index*

The speed reduction index reflects the ratio of the relative speed change between congested and free-flow conditions. Congestion usually occurs when the index exceeds 4 to 5 [2]. This concept provides a value that is easily understood by all audiences (nontechnical and technical), and its continuous scale (with numerical values between 0 and 10) offers more information on the magnitude of congestion in severely congested operating conditions than the LOS concept.

$$\text{Speed reduction index} = [1 - (\text{actual travel speed} / \text{free-flow travel speed})] \times 10 \quad (1)$$

b) *Congestion index*

The congestion index was developed by D'Este et al. and Taylor [1], and is computed as follows:

$$\text{Congestion index} = [(\text{actual travel time}) - (\text{free-flow travel time})] / [\text{free-flow travel time}] \quad (2)$$

TABLE I. PEAK HOUR AND DELAY PER ROUTE PER YEAR FOR A TYPICAL AND NON-TYPICAL DAY

Peak Hour and Delay per Route per Year	Year											
	2011		2012*		2013		2013 (incl. Fri)		2013 (holiday)		2014	
	Peak Hour	Delay (min)	Peak Hour	Delay (min)	Peak Hour	Delay (min)	Peak Hour	Delay (min)	Peak Hour	Delay (min)	Peak Hour	Delay (min)
R304 Bottelary/Bird	7-8am	7.83	7-8am	6.94	7-8am	10.81	7-8am	10.48	7-8am	2.60	7-8am	12.01
R304 Bird/Bottelary	5-6pm	2.45	-	-	5-6pm	3.29	5-6pm	3.30	5-6pm	3.14	5-6pm	6.34
R44 Kromme Rhee/Bird	7-8am	4.02	7-8am	4.17	7-8am	6.17	7-8am	6.12	7-8am	4.47	7-8am	9.96
R44 Bird/Kromme Rhee	7-8am	1.85	-	-	5-6pm	1.77	-	-	-	-	7-8am	1.41
R44 Annandale/Van Reede	7-8am	7.90	7-8am	8.10	7-8am	8.86	7-8am	8.76	4-5pm	1.43	7-8am	8.09
R44 Van Reede/Annandale	5-6pm	3.74	-	-	5-6pm	4.31	5-6pm	4.28	4-5pm	1.06	5-6pm	4.87
R310 Baden Powell/Strand	7-8am	4.85	7-8am	5.36	7-8am	6.48	7-8am	6.41	5-6pm	2.02	8-9am	2.13
R310 Strand/Baden Powell	4-5pm	1.55	-	-	4-5pm	1.35	-	-	-	-	5-6pm	9.88
Adam Tas/Piet Retief via Dorp	7-8am	2.83	-	-	4-5pm	2.70	4-5pm	2.53	4-5pm	1.26	7-8am	2.53
Piet Retief/Adam Tas via Dorp	4-5pm	7.02	-	-	1-3pm	3.37	1-3pm	3.45	7-8am	2.01	4-5pm	4.97
Bird/Van Reede via Piet Retief	5-6pm	8.74	-	-	5-6pm	7.04	5-6pm	7.09	4-5pm	2.10	7-8am	8.04
Van Reede/Bird via Piet Retief	5-6pm	8.80	-	-	5-6pm	9.26	5-6pm	9.17	5-6pm	8.33	5-6pm	9.33
Bird/Van Reede via R44	5-6pm	12.79	-	-	5-6pm	6.54	5-6pm	6.56	7-8am	6.81	4-5pm	6.45
Van Reede/Bird via R44	5-6pm	8.50	-	-	7-8am	12.92	7-8am	12.73	7-8am	5.87	5-6pm	6.51

*For 2012, only the peak-hour congestion of each arterial was studied.

A value of 0 indicates a very low level of congestion, as the travel condition is close to the free-flow condition in this case [1]. A value greater than 2 corresponds to a very congested condition [1].

c) Travel rate

Travel rate is the rate of motion, in min/km, for a specified roadway segment or trip. It is the inverse of speed and is calculated by dividing the segment travel time (min) by the segment length (km):

$$\text{Travel rate} = \text{travel time} / \text{segment length} \quad (3)$$

d) Delay rate

Delay rate is the rate of time loss for vehicles operating in congested conditions, in min/km, for a specified roadway segment or trip. It is calculated as the difference between the actual travel rate and the acceptable travel rate. Literature suggests that acceptable congestion standards may be related to congestion perceived by travellers. Motorists are usually aware of congestion when travel speeds reduce to 60 to 70 % of the free-flow speeds. This theory was adopted in this paper, applying an awareness at 70%.

$$\text{Delay rate} = (\text{actual travel rate}) - (\text{acceptable travel rate}) \quad (4)$$

e) Relative Delay Rate

Relative delay rate is a dimensionless measure that is used in this paper to compare the relative congestion on the various selected routes. It is calculated as the delay rate divided by the acceptable travel rate.

$$\text{Relative delay rate} = (\text{delay rate}) / (\text{acceptable travel rate}) \quad (5)$$

f) Delay ratio

Delay ratio is a dimensionless measure also used to compare the relative congestion levels on the various selected routes. It is calculated as the delay rate divided by the actual travel rate.

$$\text{Delay ratio} = (\text{delay rate}) / (\text{actual travel rate}) \quad (6)$$

IV. RESULTS

The results of the applied congestion indices and measures for a typical day are presented in *Addendum A*. The greater the value, the more severe the congestion. Negative values result when the actual travel conditions are better than the acceptable travel conditions. After computing the arterial speed reduction and congestion indices, it was apparent that the outbound and inbound arterials experience little congestion in the morning and afternoon, respectively. The remaining congestion measures were thus not applied to these routes.

The results of route Van Reede/Bird via Piet Retief for 2014 are typed in italics and underlined, as there must be an error in the obtained free-flow data. The sample size was less than 10, which possibly explains this error.

V. DISCUSSION OF RESULTS

Beginning with the speed reduction index, all morning values have been above 4 since 2011, except for the previously-mentioned outbound arterials. In 2014, they are all above 5 in fact, with the exception of the R310 Strand/Baden Powell route. There was a construction/maintenance zone on this route at the time, which influenced the data. Noteworthy is however the impact this zone had on the outbound afternoon traffic. The negative delay rates, relative delay rates and delay ratios of 2011 and 2013, amplified to values above 0 in 2014. Overall, the afternoon arterial traffic conditions were all below 4 in 2011 and 2013, but increased slightly above 4 on the R304 Bird/Bottelary and R44 Van Reede/Annandale arterial routes. Opposed to the R310 Strand/Baden Powell route, there are not any other known factors that could account for only a temporary increase in congestion for any of the other studied routes (e.g. long-lasting adverse weather, special events, major accidents, etc.).

The congestion index of the more congested routes lies around 1.3 and 1.7 for 2013 and 2014. This is an increase from 2011, where almost all values lay below 1.4. The afternoon congestion of the segment of R44 in town (both directions), however, encountered its worst congestion in 2011, with improvements visible since then. This is substantiated by all congestion measures applied to the probe data. These improvements are most likely not explained by less motor vehicles, but rather efficiency improvements of the traffic signals. This route nevertheless remains amongst the most congested routes in Stellenbosch.

The slowest average travel speeds (highest travel rate) are currently encountered on the studied segment of Dorp Street (both directions) and the routes Bird/Van Reede and vice versa along Piet Retief.

Surprisingly, the comparison of a typical-day traffic to Friday-traffic showed little dissimilarity. On the arterials, peak delay on Fridays (am and pm) differs only slightly to typical-day peak delay. In truth, it is fractions of a minute less. The opposite was observed for the interior roads of Stellenbosch.

The holiday period results in a shift of the peak hour for some routes. Inbound arterials experience far less morning congestion during this time, with a vast decrease in delay occurring on the two 'problem' arterials: R304 Bottelary/Bird and R44 Annandale/Van Reede. In-town congestion decreases slightly for most routes.

VI. CONCLUSION

This paper aimed to quantify Stellenbosch, South Africa, congestion beyond traffic volume counts and personal perception.

To conclude, the current traffic condition in Stellenbosch, gives reason for concern. There are too many vehicles on the

extended Stellenbosch road network at specific hours of the day.

The growth of congestion (since 2011) is inconsistent, but present (e.g. inbound peak delay on R44 Kromme Rhee/Bird increased by just over 60% from 2013 to 2014, and this route has become the most congested arterial in the morning). The other two heavily congested inbound arterials (R304 Bottelary/Bird and R44 Annandale/Van Reede) share similar morning congestion levels, but their afternoon outbound congestion has not only intensified over the years, but is almost twice that of R44 Bird/Kromme Rhee.

For 2014, the studied town-outbound segment of Dorp Street (Piet Retief/Adam Tas) has the most severe peak-hour congestion of all the studied routes. There are no alternative routes for its users, as all alternative routes in some way lead to those routes next on the list of most congested routes, for the same time period.

To generalise, the level of congestion on the arterials is worse in the morning (compared to the afternoon), but in-town adverse congestion is variable, tending to occur slightly more in the afternoon, however.

This study has verified the fact that the university and school traffic greatly contributes to the overall traffic-congestion problem in Stellenbosch, as holiday-time inbound morning arterial travel times are on average 54% that of term-time travel times.

The historical nature of Stellenbosch, its prominent aesthetic value and insufficient open land in the CBD, constrain the expansion of the existing road network. Solutions to the problem are thus limited to optimising the efficiency of the current system, but more importantly the search for alternative-mode transport systems (e.g. Park-and-Ride and Bus Rapid Transit schemes).

VII. FURTHER RESEARCH

The use of (on-board) probe data is not entirely without cons. Probe data does not reflect on the vehicle type or trip purpose. The analysis of each of their contributions to the congestion should be performed, so that various focus groups can be identified. Solutions for these focus groups should be proposed and the benefits of these solutions assessed.

REFERENCES

- [1] K. Hamad, and S. Kikuchi "Paper no. 02-2770: developing a measure of traffic congestion – fuzzy inference approach," TRB, November 2002.
- [2] T. Lomax et al., "NCHRP report 398: Quantifying congestion: volume 1 - final report," TRB, National Research Council, Washington D.C., 1997.
- [3] Statistics South Africa, 2011, Stellenbosch [Online], Available: http://beta2.statssa.gov.za/?page_id=993&id=stellenbosch-municipality [2014, August 18].
- [4] Stellenbosch University, Stats for 2014 [Online], Available: <http://www.sun.ac.za/english/lists/stats%20for%202014/tiles.aspx> [2014, October 21].
- [5] Vela VKE Engineers, "Stellenbosch Municipality: comprehensive integrated transport plan", 2nd ed., March 2011, chapter 5, pp.8-18.

ADDENDUM A. CONGESTION MEASUREMENT RESULTS FOR A TYPICAL DAY

Route	Year	Speed Reduction Index		Congestion Index		Travel Rate		Delay Rate		Relative Delay Rate		Delay Ratio	
		<i>am</i> ^a	<i>pm</i> ^b	<i>am</i>	<i>pm</i>	<i>am</i>	<i>pm</i>	<i>am</i>	<i>pm</i>	<i>am</i>	<i>pm</i>	<i>am</i>	<i>pm</i>
R304 Bottelary/Bird	2011	5.71	2.44	1.33	0.32	1.99	-	0.77	-	0.63	-	0.39	-
	2012	5.11	-	1.05	-	1.97	-	0.60	-	0.43	-	0.30	-
	2013	6.32	2.60	1.71	0.35	2.49	-	1.18	-	0.90	-	0.47	-
	2014	6.17	1.41	1.61	0.16	2.83	-	1.28	-	0.83	-	0.45	-
R304 Bird/Bottelary	2011	1.86	2.52	0.23	0.34	-	1.15	-	-0.08	-	-0.06	-	-0.07
	2013	1.33	2.97	0.15	0.43	-	1.33	-	-0.01	-	-0.004	-	-0.004
	2014	1.55	4.21	0.18	0.76	-	1.57	-	0.27	-	0.21	-	0.17
R44 Kromme Rhee/Bird	2011	4.12	1.68	0.70	0.20	1.39	-	0.22	-	0.19	-	0.16	-
	2012	4.32	-	0.76	-	1.37	-	0.26	-	0.23	-	0.19	-
	2013	5.14	1.61	1.06	0.19	1.70	-	0.52	-	0.44	-	0.31	-
	2014	6.52	2.54	1.88	0.34	2.17	-	1.09	-	1.01	-	0.50	-
R44 Bird/Kromme Rhee	2011	2.53	2.27	0.34	0.29	-	1.00	-	-0.10	-	-0.09	-	-0.10
	2013	2.21	2.04	0.28	0.26	-	1.04	-	-0.14	-	-0.12	-	-0.14
	2014	1.91	1.62	0.24	0.19	-	1.01	-	-0.20	-	-0.16	-	-0.20
R44 Annandale/Van Reede	2011	5.87	2.08	1.42	0.26	2.12	-	0.87	-	0.70	-	0.41	-
	2012	6.14	-	1.59	-	2.08	-	0.93	-	0.81	-	0.45	-
	2013	6.30	3.06	1.71	0.45	2.21	-	1.04	-	0.89	-	0.47	-
	2014	6.12	2.94	1.58	0.42	2.08	-	0.93	-	0.80	-	0.45	-
R44 Van Reede/Annandale	2011	2.24	3.41	0.29	0.53	-	1.34	-	0.08	-	0.06	-	0.06
	2013	2.73	3.70	0.38	0.60	-	1.41	-	0.14	-	0.11	-	0.10
	2014	3.21	4.32	0.47	0.77	-	1.48	-	0.28	-	0.23	-	0.19
R310 Baden Powell/Strand	2011	5.37	2.31	1.16	0.30	2.11	-	0.71	-	0.51	-	0.34	-
	2012	5.61	-	1.28	-	2.23	-	0.83	-	0.60	-	0.37	-
	2013	5.97	2.85	1.48	0.40	2.53	-	1.07	-	0.74	-	0.42	-
	2014	3.15	3.04	0.46	0.44	1.38	-	0.03	-	0.02	-	0.02	-
R310 Strand/Baden Powell	2011	2.57	2.65	0.35	0.36	-	1.30	-	-0.07	-	-0.05	-	-0.05
	2013	2.48	2.56	0.33	0.35	-	1.25	-	-0.08	-	-0.06	-	-0.06
	2014	2.72	6.05	0.37	1.76	-	2.35	-	1.02	-	0.77	-	0.44
Adam Tas/Piet Retief via Dorp	2011	5.83	4.96	1.40	0.99	4.71	3.90	1.90	1.09	0.68	0.39	0.40	0.28
	2013	4.85	5.64	0.94	1.30	3.54	4.19	0.94	1.58	0.36	0.61	0.26	0.38
	2014	5.90	5.62	1.44	1.28	4.16	3.88	1.72	1.45	0.71	0.60	0.41	0.37
Piet Retief/Adam Tas via Dorp	2011	5.61	7.86	1.28	3.68	4.16	8.55	1.55	5.94	0.59	2.28	0.37	0.69
	2013	4.14	5.67	0.71	1.32	3.14	4.25	0.51	1.62	0.19	0.62	0.16	0.38
	2014	5.31	7.26	1.13	2.65	3.79	6.49	1.25	3.95	0.49	1.56	0.33	0.61
Bird/Van Reede via Piet Retief	2011	3.70	5.07	0.70	1.19	3.44	4.40	0.34	1.30	0.11	0.42	0.10	0.30
	2013	4.96	5.25	0.99	1.11	3.74	3.97	1.05	1.28	0.39	0.47	0.28	0.32
	2014	5.12	4.12	1.05	0.70	4.75	3.95	1.44	0.63	0.43	0.19	0.30	0.16
Van Reede/Bird via Piet Retief	2011	4.35	5.05	0.77	1.04	3.82	4.36	0.74	1.28	0.24	0.41	0.19	0.29
	2013	6.08	5.95	1.55	1.47	4.69	4.53	2.06	1.91	0.79	0.73	0.44	0.42
	2014	<u>-6.52</u>	<u>-3.66</u>	<u>-0.39</u>	<u>-0.27</u>	<u>3.82</u>	<u>4.62</u>	<u>-5.20</u>	<u>-4.40</u>	<u>-0.58</u>	<u>-0.49</u>	<u>-1.36</u>	<u>-0.95</u>

Bird/Van Reede via R44	2011	6.55	7.57	1.90	3.21	3.55	5.04	1.80	3.29	1.03	1.88	0.51	0.65
	2013	6.27	6.48	1.68	1.84	3.46	3.67	1.62	1.83	0.88	0.99	0.47	0.50
	2014	6.30	6.51	1.70	1.86	3.42	3.63	1.61	1.82	0.89	1.00	0.47	0.50
Van Reede/Bird via R44	2011	6.55	6.83	1.90	2.23	3.27	3.56	1.66	1.95	1.03	1.21	0.51	0.55
	2013	7.80	6.26	3.54	1.74	6.06	3.57	4.15	1.67	2.18	0.87	0.69	0.47
	2014	6.08	6.24	1.55	1.71	2.93	3.06	1.29	1.42	0.78	0.86	0.44	0.46

^a. 7-8am.

^b. average of 4-5pm and 5-6pm.

Performance Evaluation of a Low Cost Vision-Based Traffic Flow Monitoring System

Rose Nakibuule
Makerere University
Kampala, Uganda
rnakibuule@cis.mak.ac.ug

John Quinn
Makerere University
Kampala, Uganda
jqinn@cit.ac.ug

Abstract—Traffic flow monitoring systems aim to provide accurate, complete and timely data for effective management of congestion. Conventional traffic flow monitoring systems are not suitable for deployment in crowded cities in the developing-world, due to the often chaotic nature of traffic there and typically limited budgets. In this paper we present the design and implementation of a low cost vision-based traffic flow monitoring system for such contexts using a mobile phone camera as a unit for data capture and transmission. We describe its architecture and hardware platform as well as the procedure used in traffic flow speed estimation. We also evaluate the system with performance using field experiments in Kampala, Uganda. From the experiments, we observe that the system provides a feasible method of continuously monitoring traffic congestion while reducing deployment costs drastically compared to other technologies in current use. The system also performs well according to other criteria such as ability to operate without maintenance.

I. INTRODUCTION

Many cities in the developing world are experiencing problems of increased traffic and congestion as a result of rapidly increasing population. For example population in Kampala has increased by about 600% and the population of vehicles by about 1400% between 1948 and 2012 [1]. As a result traffic and transportation managers have been challenged with how to manage the increasing traffic flows and congestion, as well as how to improve roadway capacity and efficiency with limited budgets and space for expansion. The problem is compounded by lack of real-time information on traffic flow, congestion levels and lack of resources for optimizing these levels. Lack of information is particularly problematic because traffic flow and congestion levels exhibit spatial and temporal unpredictability. They are spatially unpredictable in that we can find two parallel roads when one is heavily congested while the other is almost empty. In the temporal sense, heavy traffic flow and congestion levels occur at unpredictable times such that traffic flow and transportation managers as well as road users can not predict the traffic flow and congestion levels until they occur.

Although there exist a variety of solutions for automatic traffic flow and congestion monitoring, these solutions are not suitable for chaotic traffic flows which characterize traffic found in crowded cities in developing-world [2], as well as having prohibitively high costs in terms of purchase, installation and maintenance [3]. In an endeavor to alleviate the problems of traffic flow, transportation managers in developing-world cities have resorted to use of closed-circuit television camera (CCTV) networks to monitor traffic flow and

congestion levels. This has been hindered by high costs of CCTV cameras, costs of installation and maintenance [4], [5]. In addition, they lack sufficient staff to process and analyze the data from the CCTV networks and sufficient knowledge to maintain the networks. For example in Kampala about 50 CCTV cameras were installed in 2007 but traffic flow management is still a problem due to lack of sufficient staff to process and analyze the data. At the time of writing, many of the installed CCTV cameras are currently nonfunctional due to inadequate maintenance.

In this paper we present a prototype for vision-based traffic flow monitoring which employs a mobile phone camera as a unit for data capture and transmission. The paper expands on our previous work by providing field experiment results for assessing accuracy of the prototype in terms of traffic flow speed estimation, prototype cost (assembling and installation), its deployment feasibility and maintenance.

II. RELATED WORK

Several attempts to design and implement systems for chaotic traffic flow monitoring have been made. The use of cellular phones as probes to monitor traffic flow and congestion have been reported in [6], [7]. Cellular phones probes provide rich information on traffic flow but raises a lot privacy concerns from the participants since they may end-up revealing their private information [8], [9]. Another approach is the use of acoustic sensors placed at known distances between them [10], [9], [2]. In [2], Sen *et al*; proposes an algorithm for estimating traffic speed based on Doppler effect of the sound originating from vehicular horns. The algorithm provides 70% accuracy, but incurs high computational costs [9]. In order to reduce on the computational cost of the algorithm, Sen *et al* [9] uses the RF link variations to estimate the congestion level (traffic queue length). The algorithm achieves a 90% accuracy in congestion level prediction but requires a lot of data for training[11] at the same time do not provide visual information about traffic scenes.

There have been several attempts to monitor traffic from images and videos of chaotic traffic scenes [11], [12], [13], [14], [15]. The work by Idé *et al* [12] and Jain *et al* [13] noise data from resolution camera to monitor traffic congestion. The research in [13] relies on the usage of image feeds from CCTV camera which are costly for the developing- world while work in [12] still lacks empirical evaluation of the accuracy of the algorithms yet. Sen *et al* [11] describes a method for evaluating a road traffic congestion prediction algorithm based on colored

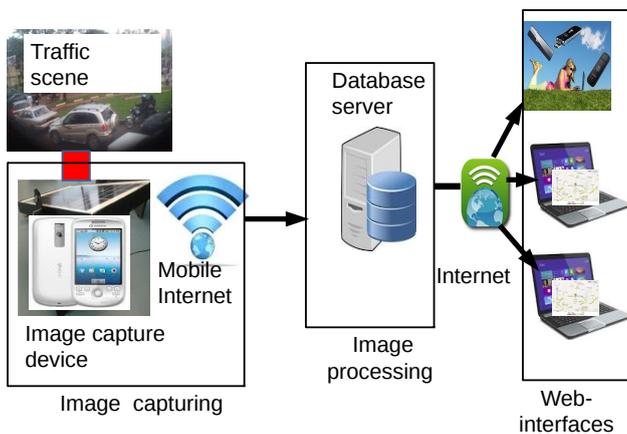


Fig. 1. Architecture of traffic monitoring system.

strips placed on the road side. The algorithm operate by calculating the percentage of the strip visible in the image through image segmentation which makes it not suitable for cluttered chaotic traffic where image segmentation is hard to archive at the same time employs high resolution expensive cameras.

Due to the limitations in current algorithms, in this paper we present a technique to measure speed based on video processing of chaotic traffic and its evaluations. The techniques does not perform any background subtraction/segmentation (which is often impractical on very crowded and cluttered road views), but rather the speed computations use feature correspondences between consecutive frames.

III. PROTOTYPE DESIGN, IMPLEMENTATION AND TESTING

The prototype consists of three components that is the image capture, image processing and a set of web-interfaces (Fig 1).

A. Image capturing Component

The image capturing component is used to capture images from traffic scene. It consists of an Android Ideos U8150 mobile phone, a 7.2Ah battery pack, a 14W, 22V solar panel, a charging regulator and a steel box. The camera on the mobile phone is programmed to capture a predefined number of images at predefined time interval and upload them to the server through a wireless internet connection. The communication between the phone and the server is done through the Advanced Message Queuing Protocol (AMQP) which is an open standard application layer protocol for message-oriented middleware. By using this protocol we are able to send multiple image files at once and hence reducing on the bandwidth needed in uploading files. To test the feasibility of this mechanism we performed field experiments where monthly subscription for 350 MB of data at a cost of 15000 Uganda shilling (\$6) was made at the start of the month of October 2012. This bandwidth was able to sustain the Internet connection for the phone for the whole month (between 1st October and 31st



Fig. 2. Image capturing unit. Top left: internal components; top right: assembled unit; bottom left: wiring for external aerial; bottom right: deployed unit.

October 2012). The amount of data captured and sent to the server by the camera set to capture 5 images each of 1280 x 768 pixels at 2 minute interval for 24 hrs and upload interval of 2 in that month was 30GB compared to 350MB used in the upload.

The solar panel is used to charge the unit and its installed on the top of a steel box which encloses the mobile phone, the battery and the regulator. The solar panel tops up the battery pack via the charging regulator so that the unit has extra charge in case of several consecutive overcast. The mobile phone is charged by the battery pack. The arm extending from the solar panel allows the unit to be bolted to a wall or post, and the camera can be rotated through two axes. The steel box offers good protection to the mobile phone, the battery and the regulator from bad weather conditions and theft but it acts a Faraday cage for the mobile phone by cutting out reception for the mobile phone. In order to have continuous reception, we connect a wire from the mobile phone's internal antennae to a wire outside and we drill small holes in the bottom of the steel box to allow fresh air to circulate in the system so as the unit does not over heat. For a complete assembly of the various parts of the image capture unit see Fig 2.

B. Image processing component

The image processing component runs on the server, and comprises of several steps.

1) Road geometry estimation and camera calibration:

Camera calibration is used to determine the projection equation between the world coordinate and image geometry. i.e to map a point $(x^{(im)}, y^{(im)})$ in the image plane to a point $(x^{(w)}, y^{(w)})$ in the world plane. In this work we perform an off-line manual calibration based on direct linear transformation (DLT) analysis as in [16]. A square grid measuring 1m x 1m is placed in the world coordinated system of a traffic scene with points at the border forming the set of control points denoted

by $\left\{ \left(x_i^{(w)}, y_i^{(w)} \right) \mid i = 1, \dots, 4 \right\}$ are obtained using GPS system and their corresponding image coordinates denoted by $\left\{ \left(x_i^{(im)}, y_i^{(im)} \right) \mid i = 1, \dots, 4 \right\}$ are obtained from the captured images. Given this correspondence we compute the perspective transform \mathbf{P} between the image plane and the World plane as portrayed in equation 1 through direct linear transformation analysis and we can infer two this transformation whenever we want to obtain the "real world" speed on the traffic

$$\begin{bmatrix} \lambda * x_i^w \\ \lambda * y_i^w \\ \lambda \end{bmatrix} = P \begin{bmatrix} x_i^{im} \\ y_i^{im} \\ 1 \end{bmatrix} \quad (1)$$

where P is a 3×3 matrix, $i = 1, 2, \dots, N$ with $N = 4$. This process is performed every time the camera changes position or location.

2) **Specifying regions of interest (ROIs):** We define a polygon R in the image enclosed by the image coordinates $\left\{ \left(x_i^{(im)}, y_i^{(im)} \right) \mid i = 1, 2, \dots, N \right\}$ where $N \geq 4$ specifying the region of interest we want to monitor and speed estimations are based on this region of interest. This is done manually at the beginning of the monitoring process.

3) **Feature flow extraction:** Informative features are extracted from the uploaded images using the method described in [14]. Given an input of images taken at regular intervals; the first step is to take pairs of images and calculate the correspondence between the two, so that we can obtain flow vectors corresponding to every moving object. In this work we use the feature flow approach to extract the moving objects in that it is first and it can be incorporated into object recognition tasks. To extract important features or key point features, we use the scale and rotation-invariant detector and descriptor called Speeded-Up Robust Features (SURF) developed by Bay *et al* [17] in OpenCV, being faster than the popular SIFT descriptor [18]. From the extracted SURF features, we establish correspondences between correspondences between frames. SURF features are vectors which describe a visual feature in a representation invariant to rotation and scale. The SURF features are identified by applying difference of Gaussian convolutions of second order derivatives to an image at different scales. The descriptors are then created based on Sum of Haar Wavelet Responses around the interesting point. Then we use the SURF features to calculate correspondences between each frame, giving us a set of motion vectors /flow vector in the coordinates of the image as in [14].

We then infer road geometry in relation to the camera, to project those vectors into 'real-world' coordinates, allowing us to calculate speeds in km/h.

C. Web interfaces

The web-interfaces provides a mechanism for visualizing traffic information and for monitoring and updating the camera. Two kinds of web-interfaces are implemented.

1) **Administrative web-interfaces:** These interfaces provides a way for the administrator to monitor and update information about the installed cameras. Through these interfaces (s)he is able change camera setting which may include, the number of images to be captured, the image capture and upload interval.

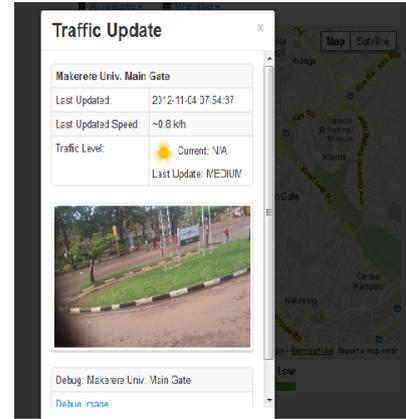


Fig. 3. Information visualization

2) **Information display web-interface:** This interface is used to display traffic information to the road user. Fig 3 shows a sample interface for displaying traffic information where the user can view current level of traffic flow at the same time visualize it from the image itself originating for the scene.

IV. SPEED ESTIMATION

During the estimation of traffic flow speed, we first specify the direction we want to monitor traffic. This is necessary for instance where there are two traffic flows moving at relatively the same speed in the opposite direction; without constraining the calculation to particular regions of interest and directions of flow may end up canceling one another and we observe zero flow. By specifying the direction of flow the application will consider only flow vectors moving in that particular direction in calculating the traffic speed within the region of interest.

A. Motion vector filtering by direction

Since the traffic flows we are handling can move in any direction with no defined lane markings, we first define a direction vector $u(\delta x, \delta y)$ which specifies the direction in which we want to monitor traffic. Then for each flow vector $\{x_i^w \mid i = 1, 2, \dots, N\}$ computed, we compute the angle α between the flow vector as

$$\alpha = \arccos \left(\frac{u * x_i^w}{|u| * |x_i^w|} \right), \quad (2)$$

and set a threshold T such that a flow is said to be moving in the selected direction if $\alpha \leq T$. After obtaining all the flows moving in the specified direction, we apply a set of filtering rules to filter out all flows in the direction which may come as a result of mismatches.

1) **Filtering fast moving motion vectors:** When the selected motion vectors are mapped into the world coordinate system, some flows are moving very fast, so we filter out all the flows which are moving with speed more than 100km/h as they are most likely not vehicle flows; since most of the vehicles in the free flow tend to move with an average speed of between 60km/h and 70km/h as presented in [19].

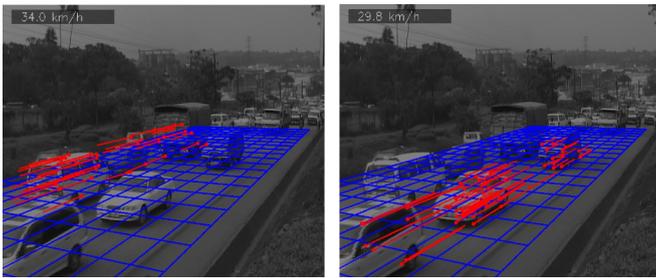


Fig. 4. Filtering motion vectors.

TABLE I. COST OF HARDWARE

Component	Cost (\$)
Mobile phone	76.9
Battery pack	30.7
Regulator	7.7
Solar panel	19.2
frames,bolts and labor	19.2
Steel box	19.2
Installation	7.7
Total	180.6

2) *Filtering motion based on (meter/pixel) ratio:* Here we investigate the ratio ρ defined by $\frac{L_{meter}}{L_{pixel}}$ where L_{meter} is the length of flow in meters and L_{pixel} is length of flow in pixels. When ρ is high beyond a certain threshold T_{ratio} , we filter out the flow since this could indicate a mismatch between feature flows.

V. PERFORMANCE EVALUATION

The section presents results of field experiments carried out to access the accuracy our prototype in traffic flow speed estimation and costs for the hardware.

A. Hardware (Image capturing component)

1) *Cost:* The cost of the various components needed in the construction and installation of the image capturing unit is given in table I. The total cost of unit with installation costs inclusive is 180.6 USD compared to the cost of Traffic CCTV and its installation which is estimated between 9,000 and 19,000 USD [4], [5], while a machine vision sensor at an intersection is estimated between 16,000 and 25,500 USD per installation [3]; which makes it a more cost effective solution option for developing-world.

2) *Robustness to changes in weather conditions:* To test the robustness of the unit to the changes to weather condition the unit was installed in the field (Makerere main gate) for a period of 22 months from September 2012 to July 2014. The unit was exposed to various weather conditions and remained functional. Fig 5 shows the internal structure of the unit and solar after deployment.

3) *Maintenance:* As we tested the unit robustness to changes in weather conditions; we also tested how frequent the unit parts needed maintenance or replacement. It was observed throughout the period that all parts functioned as required and no maintenance or replacements were required, despite the unit being fully exposed to tropical weather throughout the testing period.



Fig. 5. Internal parts of the unit and solar panel after 22 months of continuous deployment.

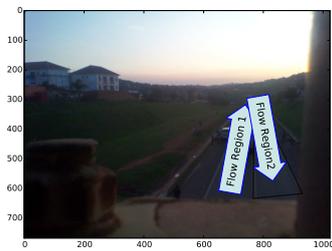
B. Software

Software evaluation was done to assess how accurate it is in estimation of traffic flow speed. The system was tested with live traffic from the field for 3 hour. The system was set to capture five images every after 2 seconds delay and upload them at 2 second interval to the server. As the images arrived at the server the image processing software running at the server is triggered to extract the average flow speeds which from the uploaded images. Then the estimated speed is used in calculation of a speed moving averages as our long term is to estimate the average speed of traffic flow within a given time frame. A car with a GPS logger system was drove through the camera field of view. The speed estimates from the GPS logger was used to provide ground truth information on the accuracy of the software in speed estimation. The results in Fig 6 shows the speed moving average for both data obtained from the GPS logger and the software estimation with a window of 10 seconds. From the results it is observed that:

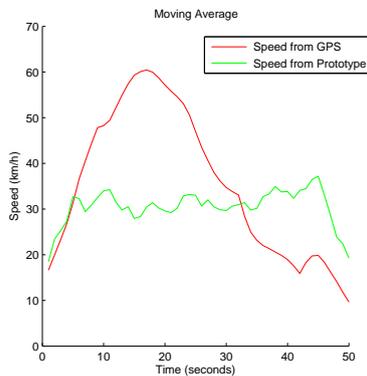
- 1) When the car is within the region of interest approximations from the GPS and software are similar. In figure 6(b)) the car started in the region of interest 1 and moved past the region; while in figure 6(c)) the car started away from the region and moved towards the region of interest and beyond. From 6(b)) we see the results from both the GPS and software are similar between the first 10 seconds the time when the car was in the region of interest and the same applies to Fig 6(c)) when the car approached the RIO between time interval of 50 and 60 seconds.
- 2) As the car moves out of the region of interest there is a lot of variations in both results. This is because, the status of traffic flow in the current location of the car may differ from the one in the region of interest. For example in the current location of the car the traffic flow may be free flow while in the region of interest, there is no flow (empty).
- 3) When the region of interest selected is large, the average flow speed becomes small see Fig 6(d). This because it includes a large set of objects/vehicles whose variations in speed becomes significant.

VI. CONCLUSION

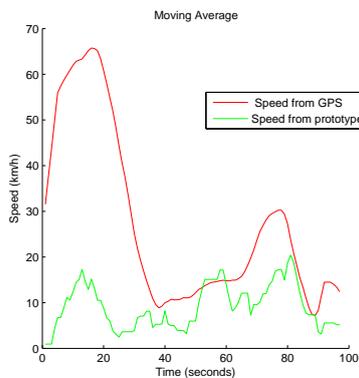
The study has explored the evaluation of the performance of a software and hardware prototype for video-based traffic flow monitoring using the camera of a mobile phone as the basic unit of data capture and transmission. The evaluations



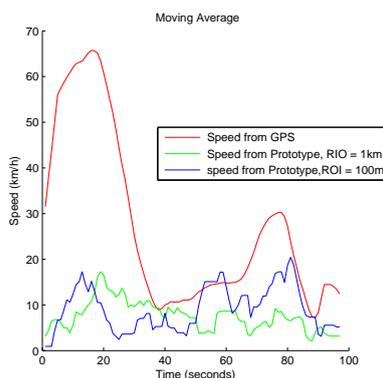
(a) Separate regions of interest within a single field of view.



(b) Speed in region 1



(c) Speed in region 2



(d) Average speed vs size of ROI

Fig. 6. Comparisons of inferred speed from vision system with speed recorded by GPS logger on board a vehicle.

shows that the prototype reduces the cost of deployment by 82% compared to video systems currently on the market at the same time providing public real time access to data from permanent installations of our prototype in Kampala as a resource for researchers of developing-world.

The initial results obtained shows that the performance of the software varied depending on the size/length of the region of interest but still suitable for crowded cities in that it makes no assumption of vehicles travelling in fixed lanes, absence of clutter or possibility of segmenting individual vehicles.

The main limitation in the software evaluation was the inability to get enough vehicle probes equipped with GPS in the regions of interest to provide enough ground truth information which could help in improving the evaluation results of the software. The other challenge faced was determining the appropriate interval for capturing images since with large time interval, a lot of information is lost especially during free flow while with short intervals lot of data is sent on the network which leads to congestion hence reducing the performance of the system. This calls for further investigation for better understanding of the appropriate interval and if possible setting dynamically depending on the level of traffic flow.

REFERENCES

- [1] P. S. (POPSEC), "The state of uganda population report 2012; uganda at 50 years: Population and service delivery; challenges, opportunities and prospects," 2012. [Online]. Available: <http://popsec.org/wp-content/uploads/2013/10/SUPRE-REPORT-2013.pdf>
- [2] R. Sen, B. Raman, and P. Sharma, "Horn-ok-please," in *Proceedings of the 8th international conference on Mobile systems, applications, and services*. ACM, 2010, pp. 137–150.
- [3] G. Leduc, "Road traffic data: Collection methods and applications," *Working Papers on Energy, Transport and Climate Change*, vol. 1, p. 55, 2008.
- [4] ServieMagic.co.uk, "How much does it cost to install cctv?" 2013. [Online]. Available: <http://www.serviemagic.co.uk/tips-and-advice/how-much-does-it-cost-to-install-cctv.html>
- [5] ServieMagic, "Cctv cameras & cctv equipment price list?" 2014. [Online]. Available: <http://www.cctv-centre.co.uk/cctv/cctvprices.htm>
- [6] M. Prashanth, N. P. Venkata, and R. Ramjee, "TrafficSense: Rich monitoring of road and traffic conditions using mobile smartphones," Microsoft Research, Tech. Rep. MSR-TR-2008-59, April 2008. [Online]. Available: <http://research.microsoft.com/apps/pubs/default.aspx?id=70573>
- [7] P. Mohan, V. N. Padmanabhan, and R. Ramjee, "Nericell: rich monitoring of road and traffic conditions using mobile smartphones," in *Proceedings of the 6th ACM conference on Embedded network sensor systems*. ACM, 2008, pp. 323–336.
- [8] P. Händel, J. Ohlsson, M. Ohlsson, I. Skog, and E. Nygren, "Smartphone based measurement systems for roadvehicle traffic monitoring and usage based insurance," *IEEE Systems Journal*, 2013.
- [9] R. Sen, P. Siriah, and B. Raman, "Roadsoundsense: Acoustic sensing based road congestion monitoring in developing regions," in *Sensor, Mesh and Ad Hoc Communications and Networks (SECON), 2011 8th Annual IEEE Communications Society Conference on*. IEEE, 2011, pp. 125–133.
- [10] R. Sen, A. Maurya, B. Raman, R. Mehta, R. Kalyanaraman, N. Vankadhara, S. Roy, and P. Sharma, "Kyun queue: a sensor network system to monitor road traffic queues," in *Proceedings of the 10th ACM Conference on Embedded Network Sensor Systems*. ACM, 2012, pp. 127–140.

- [11] R. Sen, A. Cross, A. Vashistha, V. N. Padmanabhan, E. Cutrell, and W. Thies, "Accurate speed and density measurement for road traffic in india," in *Proceedings of the 3rd ACM Symposium on Computing for Development*. ACM, 2013, p. 14.
- [12] T. Idé, T. Katsuki, T. Morimura, and R. Morris, "Monitoring entire-city traffic using low-resolution web cameras," in *Proceedings of the 20th ITS World Congress, Tokyo*, 2013.
- [13] V. Jain, A. Sharma, and L. Subramanian, "Road traffic congestion in the developing world," in *Proceedings of the 2nd ACM Symposium on Computing for Development*. ACM, 2012, p. 11.
- [14] J. A. Quinn and R. Nakibuule, "Traffic flow monitoring in crowded cities," in *AAAI Spring Symposium: Artificial Intelligence for Development*, 2010.
- [15] B. Maurin, O. Masoud, and N. Papanikolopoulos, "Camera surveillance of crowded traffic scenes," in *Proc. of ITS America Twelfth Annual Meeting*, 2002, pp. 28–58.
- [16] W. xi feng, T. Yong, Z. zhong zhe, and L. hai ying, "Feasibility of using digital photography for environmental monitoring of animals in an artificial reef," in *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, ser. Part B6b, vol. XXXVII, Beijing, 2008, pp. 339–342.
- [17] H. Bay, A. Ess, T. Tuytelaars, and L. J. V. Gool, "Surf: Speeded up robust features," *Computer Vision and Image Understanding*, vol. 110, no. 3, p. 346359, 2008.
- [18] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [19] R. Nakibuule, J. Ssenyange, and J. A. Quinn, "Low cost video-based traffic congestion monitoring using phones as sensors," in *Proceedings of the 3rd ACM Symposium on Computing for Development*. ACM, 2013, p. 52.

Modelling of inter-stop minibus taxi movements: Using machine learning and network theory.

Innocent Ndibatya

School of Computing and Informatics Technology,
Makerere University, Kampala, Uganda
(Exchange student at Stellenbosch University)
Email: indibatya@cis.mak.ac.ug

MJ (Thinus) Booysen

Department of Electronics and Electrical Engineering
Faculty of Engineering
University of Stellenbosch, Stellenbosch, South Africa
Email: mjbooyesen@sun.ac.za

Abstract—Minibus taxis provide affordable alternative transport for the majority of urban working population in Sub-Saharan Africa. Often, these taxis do not follow predefined routes in their endeavours to look for passengers. Frequently, they stop by roadsides to pick up passengers and sometimes go off the main route in an attempt to fill the taxi with passengers to make the trip profitable. In addition, the destinations are changed from time to time depending on the driver. This uncoordinated movement creates a web of confusion to would-be passengers. The key aspects that are not clear to the passengers include; where to get a taxi, the waiting time and the travel time to the destination. These conditions leave taxi passengers at a very big disadvantage. In this research, we applied the concepts of machine learning and network theory to model the movements of taxis between stops. The model can be used to compute the waiting times at the stops and the travel times to a specified destination. Twelve minibus taxis were tracked for 6 months. Density-based clustering was used to discover the formal and informal taxi stops, which were modelled into a flow network with the significant stops as nodes and the frequency of departures between nodes as edges representing the strength of connectivity. A data driven model was developed. From the model, we can predict the time a passenger will have to wait at a stop in order to get a taxi and the trip duration.

I. INTRODUCTION

The medium-sized minibus taxis which carry 10 to 15 passengers dominate the public transport sector in South Africa. 60 % of South African citizens rely on them and they transport an average of 14 million people every day. However, they are not reliable and not properly regulated by government. If by chance a passenger stands at the right pick up place, they do not know how long to wait before the next taxi will pass by and the travel time to reach their destination. In our earlier study [1], we used machine learning to find the formal and informal stops (Figure 1 (a, b and c)) of taxis operating between towns in the Western Cape. During that study, it was discovered that the rate at which taxis stopped at different stops varied according to the days of the week and the times of the day. We demonstrated that by using historical GPS locations of taxis, we could predict the most probable places where a traveller could get a taxi depending on the time of the day and the day of travel. However, we were not able to tell how long the person would wait at the stop in order to get a taxi and the duration of the journey. We therefore recommended further investigations to find the waiting time at the stops, the routes taken by the taxis and the trip durations. In the current study we considered the waiting time at the stops and the trip

durations of the taxis. In this paper we model the movements of the minibus taxis between the stops in order determine (1) the waiting time at the stops, and (2) the trip durations as the taxis move from one stop to the other.

We believe that this information is key to understanding the operations of the informal public transport sector where the minibus taxis are dominant. The rest of this paper is organised as follows. Section II discusses the theoretical basis of the research, section III discusses the methods used, section IV discusses the results and section V concludes the paper. In this paper, the term *asset(s)* is used to refer to the *taxis* that were tracked during the study, and sometimes, the words are used interchangeably. The term "*modelled system*" is also often used to refer to the conceptual "experimental component set up" during the process of modelling.

II. LITERATURE REVIEW

Researchers have in the past attempted to model transportation systems and many models have been developed. These models were categorised by Zhou and Dai [2] into two, i.e Freight transportation models for goods in transit and passenger transportation models for human passengers. In developed cities, passenger transportation models have traditionally been part and parcel of urban design [3]. The main objectives of the incorporation are (1) to reduce the number of motorized trips, (2) to increase the number of non-motorized trips, and (3) to increase vehicle occupancy. Traditional transportation models have always been theoretical and based on mathematical description of processes. Such models forecast travel demand based on trip generation, trip distribution, transportation mode (public/private) and route choice [4].

The 21st century era of powerful computing introduced a new method of modelling – Data Driven Modelling (DDM). DDM uses vast amounts of data about the process and methods of computational intelligence and machine learning to model processes [5]. Machine learning deals with the construction and study of algorithms that can learn from data. Some of the methods used for machine learning include, Network analysis, Neural networks, Markov models and many more. Network analysis in particular interprets the modelled process into a network graph with discrete objects represented as vertices and the relationships between the objects represented as edges. Combined with machine learning, researchers have applied network analysis to model problems in many fields such as in neural science [6].

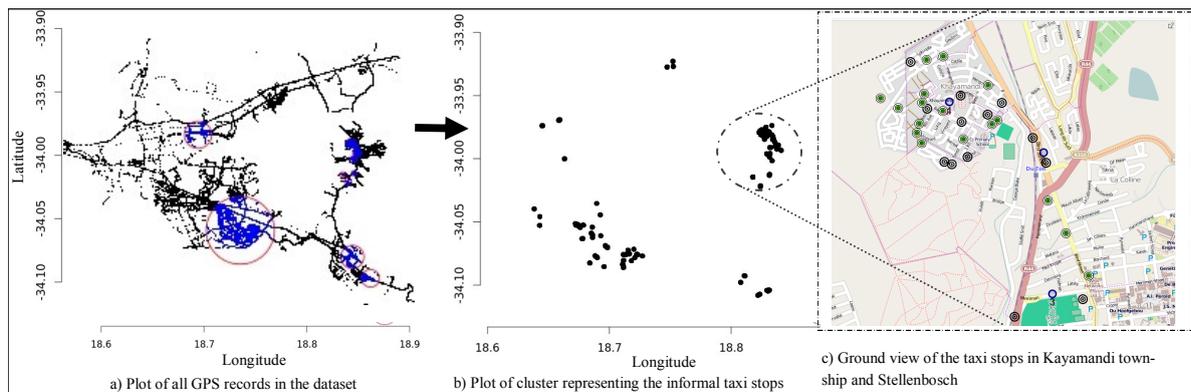


Fig. 1. Figures showing the formal and informal stops discovered during our previous study

The reviewed literature suggests that the network analysis method has successfully been used before to model processes. Furthermore, in the presence of data, DDM can be used to study processes and develop dynamic models that learn from the data. It is on this basis that we sought to apply the methods to model the movements of minibus taxis.

III. METHODS USED

In this study, we used two methods. (1) Vector overlay of GPS trajectories in a circle to map out the significant nodes/stops, and (2) network theory to model the movements of the taxis between the significant nodes/stops.

A. Data sources

Ten minibus taxis (assets) operating between Stellenbosch, Somerset West, Strand, and Bellville were equipped with GPS (Global Positioning System) tracking devices. Each device had a GSM (Global System for Mobile communications) sim card installed and would log GPS locations of the taxi to a remote server through the GSM network. Data was logged at a nominal frequency of 1Hz. Attributes of the data that was logged included - date and time, GPS location (longitude and latitude), speed and direction among others. This dataset contained a total of 1,842,570 GPS records collected over a period of 6 months (December 2013 - May 2014).

B. Vector overlay of GPS coordinates

Overlay analysis is a technique commonly used in GIS (Geographical Information Systems) studies where layers of features are analysed to find intersects, unions and clips. In this study we performed circular overlays over different regions of dense GPS points – significant clusters (Figure 2b and c). To create the overlays, centres were determined by computing the mean GPS points of eleven significant clusters discovered in the dataset when speeds were less than 2 km/h. The diameters of the overlay circles were determined by measuring the distance of the closest centres. This meant that centres that were very close to each other had smaller radii compared to those that were far apart as shown in Figure 2b. All the Western Cape GPS records (1,842,570) and eleven centres with radii ranging from 0.3 to 5 kilometres were used during the overlay analysis. After the overlay analysis, GPS intersects

were obtained (These were records intersecting the overlay circles) and labelled by appending meta-data representing the ID of the intersecting circular layers. A total of 1,573,829 intersects were obtained. These were records in the close proximity of the significant clusters since the mean point was used as the center. From this point on, two assumptions were made. (1) That each of the eleven significant clusters were major sources and destinations of taxi trips. (2) That taxis never stopped in other places during transit from one node to the other. Results from this section opened way for our next section (network flow analysis).

C. Network analysis

During this phase of our experiment, every overlaid region (significant clusters region) was represented as a node on the network (Figure 2c). The objective of this phase was to study the departure times and trip durations between the nodes for all individual dates recorded in our dataset.

An algorithm was developed that processes the movements of the taxis between the nodes. The algorithm takes the GPS intersects, the node IDs, and the periodic interval. The periodic interval is the time (in minutes) that defines the rate at which the algorithm checks on the status of an asset to determine if it has departed from the node or arrived at the node. The interval was estimated such that it was less than the expected duration of travel of an asset between any two nodes. During our experiment, we used the interval of 10 minutes as the expected minimum travel time between any two closest nodes. Figure 3 shows the sample modes of the assets at different periodic intervals. In the table, assets *A* to *L* are monitored and their modes recorded at different time intervals. For example, when 3 is recorded in a cell, it means that the asset was registered active in a cell during that periodic interval and *N* mean that the asset was recorded inactive a given periodic time.

We then analysed the records obtained in the table (Figure 3) to get a count of departures and the trip durations between nodes. During this analysis, we used a combination of two modes (active/transmitting and inactive/not transmitting) and a transition between them to explain three unique states in the modelled system – "*atNode*", "*inTransit*", "*inSleep*". An asset in our modelled system can only be in any of the three states at a time. When an asset is *active* in two

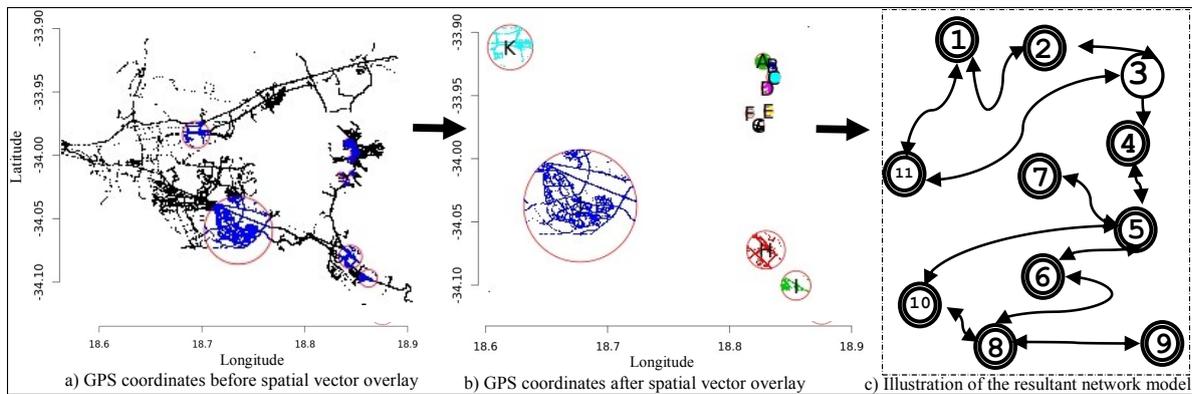


Fig. 2. Figures showing the vector overlay process and the first phase of modelling significant nodes into a graph

		Periodic interval (10 minutes)																										
		15:00	15:10	15:20	15:30	15:40	15:50	16:00	16:10	16:20	16:30	16:40	16:50	17:00	17:10	17:20	17:30	17:40	17:50	18:00	18:10	18:20	18:30	18:40	18:50	19:00	19:10	
Assets tracked	A	1	3	5	5	N	3	N	4	1	1	1	7	4	3	3	3	1	1	2	N	N	2	2	1	1	N	
	B	1	1	1	2	N	1	1	3	N	3	1	1	3	3	3	1	1	2	2	1	1	1	3	1	N	N	
	C	3	3	1	1	3	1	1	5	5	1	1	3	1	1	3	N	N	3	3	3	N	8	8	8	5	1	
	D	N	3	N	N	N	N	N	N	N	3	1	1	4	N	N	3	1	1	1	1	1	1	N	N	1	N	
	E	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	F	N	7	N	N	N	N	N	3	1	1	3	3	6	6	3	1	1	1	1	3	2	1	1	1	1	3	
	G	N	4	3	3	3	N	3	1	1	3	N	N	N	3	1	1	N	3	3	3	1	1	1	1	N	N	
	H	3	3	3	7	N	8	N	8	N	N	8	N	5	1	1	3	N	3	1	1	1	3	3	4	3	N	
	I	6	6	3	1	1	3	1	1	1	1	3	2	1	1	3	3	N	3	1	1	3	N	N	N	N	3	
	J	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N
	K	2	N	N	N	1	1	N	N	1	1	N	N	N	N	N	N	N	1	1	1	1	N	11	11	N	11	
	L	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N	N

Fig. 3. Table showing the periodic assessment of the assets (A-L) modes and the nodes(significant stops) where assets are transmitting from.

successive intervals and it is at the same node, then the asset is in "atNode" state, e.g asset B at 15:00,15:10 and 15:20. An "inTransit" state is recorded if an inactive mode(s) is observed between two active modes at different nodes, e.g asset D during the 17:00 and 17:10 time intervals. An asset is in "inSleep" state if an inactive mode is recorded between two active modes at the same node, e.g asset H during the time intervals 16:20 and 16:30.

The results of the table in Figure 3 are analysed by a separate algorithm which deduces two general matrices; (1) the count of departures from the node for every day (identified by the date), (2) the average duration between two nodes (complete trip). Essentially, the algorithm monitors the time and state of all assets for every time interval in the modelled process. It records an asset departure from one node to another if an "inTransit" state is realised (e.g in Figure 3, asset D departed from node 4 at 16:50 and arrived at node 3 at 17:20). The algorithm uses the time difference to compute and record the trip duration. For example the trip duration of asset D from node 4 to node 3 recorded between 16:50 and 17:20 was 30 minutes. Figure 4a shows the sample departure count matrix for a selected date and Figure 4b shows the sample average duration matrix for the same day in the modelled system.

IV. RESULTS

Taxis in the Western Cape tend to exhibit a stochastic behaviour. Though the routes taken by the individual taxis tend to change most of the time. It was discovered by cluster analysis that over time these routes go through some specific

places (stops) where the taxis pick up and drop passengers. However, some of these stops are not gazetted by the local authorities hence we referred to them as informal stops.

At the stops (significant), taxi departures vary according to the destination. Figure 4a shows a matrix of departure counts from different places (nodes) on a Tuesday. It is clear that there were more departures from node 3 to node 1 (177 departures) while there were no departures from node 3 to node 11. However, it was discovered that while in some cases there are no direct departures between two nodes, these nodes are still reachable through other nodes. For example, if a passenger wanted to use a taxi from node 3 (Stellenbosch) to node 11 (Bellville), he/she can get a taxi from node 3 to node 1 and then from node 1 to node 11.

By spreading the departure times at a single node throughout the day, we can tell what time segment has the most frequency of taxis departing and so we can compute the waiting time depending on the time of arrival of the traveller. Figure 4c shows the graph of varying departure frequencies from node 3 to other nodes during the day. In the plot, every line represents trend of departures to one node in the modelled system. From the graph, the peak departures from node 3 to node 1 happen very early in the morning (05:00) and at around 17:00 hours in the evening. It should be noted that the variations also depend on the destination node. For example, departures to node 6 (Somerset West) on this day have their peak at around mid-day. Therefore we can compute the waiting time at the node. Similar graphs comparing inter-node durations can be plotted and a similar analysis undertaken.

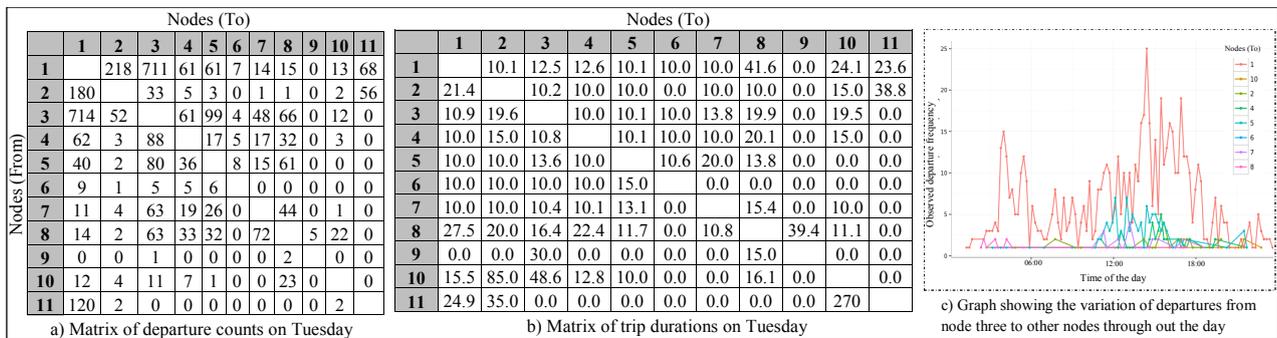


Fig. 4. Figures showing the matrices from the network modelling process and a graph of varying departure frequencies with time of the day.

Using the duration matrix, we can compute the duration of the passenger journey. For example a passenger who wishes to travel from node 3 to node 11, would take 10.9 minutes to get to node 1 and 23.6 minutes to get to node 11 hence a total of 35.5 minutes provided there is no delay at the intermediate node.

From our results, time optimisation is possible by choosing the most optimal travel choice. For example, in order to travel to node 11 from node 3, there are two choices, i.e from node 3 to 1 to 11 which takes 35.5 minutes and from node 3 to 2 to 11, which takes 58.4 minutes. Travellers can then choose the most optimal choices that suites their needs.

V. CONCLUSION

Movement of taxi between stops have been studied and modelled to provide useful information to the users of the informal public transport sector (Minibus taxi users). If utilised, users can reduce the time wasted waiting for taxis and during the journey by planning their trips in advance.

For data driven models to be more accurate and precise, there is need for a lot more data. This would assist researchers in discovering more hidden behaviour regarding the minibus taxis. Particularly there is need to study the variations of the location of stops with particular seasons of the year; the delays at the stops; and a continuous learning mechanism that will always keep the models up-to-date.

ACKNOWLEDGEMENT

The authors would like to thank Mix Telematix and Trinity Telecom for providing data for the research, Uganda Christian University for providing computing resources to test and run the experiments.

REFERENCES

[1] I. Ndiabuya, M. J. Booyesen, and J. Quinn, "An adaptive transportation prediction model for the informal public transport sector in africa," in *IEEE 17th International Conference on Intelligent Transportation Systems (ITSC)*, Qingdao, China, October 8-11 2014, pp. 2572–2577.

[2] J. Zhou and S. Dai, "Urban and metropolitan freight transportation: A quick review of existing models," *Journal of Transportation Systems Engineering and Information Technology*, vol. 12, no. 4, pp. 106 – 114, 2012. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1570667211602146>

[3] R. Cervero and K. Kockelman, "Travel demand and the 3ds: Density, diversity, and design," *Transportation Research Part D: Transport and Environment*, vol. 2, no. 3, pp. 199 – 219, 1997. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1361920997000096>

[4] H. K. transport department, "Third comprehensive transport study (cts-3)," *Journal of Urban Planning and Development*, 2011.

[5] D. P. Solomatine, *Data-Driven Modeling and Computational Intelligence Methods in Hydrology*. John Wiley and Sons, Ltd, 2006. [Online]. Available: <http://dx.doi.org/10.1002/0470848944.hsa021>

[6] H. B. Jonas Richiardi, Sophie Achard and D. V. D. Ville, "Machine learning with brain graphs," *IEEE Signal processing magazine*, Tech. Rep., 2011.

The impact of average speed over distance (ASOD) systems on speeding patterns along the R61.

N.A. Ebot Eno Akpa*, M.J. Booysen* and M. Sinclair†

Abstract—Speeding is considered to be a major contributing cause of road fatalities in Sub-Saharan Africa and South Africa in particular. The minibus taxi industry is a vibrant yet partly informal sector of public transport in South Africa, which has been associated with speed-related road fatalities. Although countermeasures have been implemented to address speeding, they have not led to significant reduction in road fatalities and adherence to legal speed limits. Among the countermeasures deployed on some highways is the Average Speed Over Distance (ASOD) system which uses cameras to enforce speed limits. In this paper, historical probe data is used to evaluate the impact of the ASOD system on speed profiles of passenger vehicles. The data also consists of speed, time and location information gathered by navigation and fleet management devices that were installed in minibus taxis. The evaluation is based on spatial differentiation (the impact on the enforcement site with ASOD versus the control site without ASOD) and time differentiation (the impact before and during ASOD enforcement). For passenger vehicles, the results show that the presence of ASOD systems caused a reduction in mean speeds and ensured compliance with speed limits at enforcement and control sites. On the other hand, the system appears to have no influence on minibus taxis, with high, yet similar average speeds measured in the enforcement and control sites during ASOD enforcement.

I. INTRODUCTION

Various speed reduction countermeasures have been used in South Africa to reduce speed-related fatalities and injuries. While some countermeasures such as rumble strips and speed humps are aimed at managing vehicle speeds, other countermeasures such as instantaneous speed cameras are aimed at enforcing compliance with posted speed limits [1]. This paper focuses on the Average Speed Over Distance (ASOD) system implemented on the R61 between Beaufort West and Aberdeen in South Africa. ASOD is a form of automated speed enforcement, which ideally promotes both speed management and compliance with posted speed limits through Average Speed Enforcement (ASE). It is usually referred to as Point-to-Point speed enforcement in parts of Europe. A number of studies have proven the effectiveness of ASODs in Australia and Europe with a significant reduction in crash rates by about 24% over a period of three years during enforcement [6][7]. The size of its effect on crash rates and speed violation may vary from region to region as [2] suggests.

The infrastructure involves the installation of Automatic Number Plate Recognition (ANPR) cameras at strategic locations along a road section. An image of the number plate is taken at the initial camera location and at any subsequent

camera locations. Image processing is used for character recognition, followed by the retrieval of vehicle data from a central database. The known distance between the cameras and the time taken to travel between both cameras is used to calculate the average speed of the vehicle along the section. A fine is issued if the calculated average speed is higher than the enforcement threshold limit of the road. Fixed or mobile cameras may be used for ASOD implementation. Most systems use fixed cameras due to lower subsequent speed violations and crash rates associated with them [2]. Camera visibility is enhanced through roadside notifications. It should be noted that ASOD as a speed enforcement countermeasure is different from instantaneous speed enforcement. While the latter has also been used to control speed [9], unlike the former, it is limited to specific locations.

A. Problem statement and Research objectives

With road transportation becoming an increasingly integral part of societal activities in Africa, the need for efficient road safety measures is growing. Speeding is often cited as the leading human factor responsible for road fatalities. Studies have shown that there is a direct relationship between vehicle speed, crash risk and crash severity [3]. According to South Africa's 2011 road traffic report, speeding contributed to about 40% of fatal crashes, due to human error [4]. As a result, modern ITS (Intelligent Transport System) safety measures such as ASODs are geared towards regulating human factors. The need to evaluate the impact of ASOD systems on speed profiles is therefore inevitable. Such an investigation needs to transcend macroscopic effects as discussed in [5] to address microscopic effects such as reductions in average speed and speed variability, while considering different modes of transport.

The objective of this paper is to investigate the effects of ASOD systems on speeding, before and during ASOD enforcement: this will be referred to as 'time differentiation' analysis. The paper also compares the effects of ASOD implementation on enforcement sites and control sites: this will be referred to as 'spatial differentiation' analysis. Unlike most previous studies, the specific objective of this paper is to investigate the impact of ASODs on private and public modes of transport. The research focuses on passenger vehicles (mainly used for private transportation) with a legal speed limit of 120 km/h, and minibus taxis (mainly used for public transportation) with a legal speed limit of 100 km/h.

II. RELATED WORK

ASE systems have been operating in certain regions for about seventeen years. The first instance was a trial form installed in 1997 in the Netherlands, which ran for five years

N.A. Ebot Eno Akpa and M.J. Booysen are with the Department of Electrical and Electronic Engineering, Stellenbosch University, e-mail: mjbooyesen@sun.ac.za

M. Sinclair is with the Department of Civil Engineering, Stellenbosch University.

before permanent implementation in 2002. In 2000, England launched its first permanent implementation after running trial versions for a year [7]. Besides South Africa, there is no documented literature on the implementation of average speed enforcement in Sub-Saharan Africa. The majority of Sub-Saharan African nations rely on police patrols, rumble strips and speed humps to control speed [10]. South Africa launched one of its first ASE technologies – known as the ASOD – in October 2011 on the R61: a 71.6 km stretch of road between Beaufort West and Aberdeen [5].

Media reports on ASE systems indicate that they have been effective in improving road safety. This is evidenced by the number of speed fines issued and reduction in road fatalities. The evaluation of the effectiveness of ASE systems is, however, a relatively new research topic. This applies particularly to the African context, where ASOD systems have been operational for less than half a decade. Hence, there is still a general lack of a credible body of research on the extent of its effects on speed management in different regions, and the availability of concrete evidence to substantiate its supposed benefits [6].

A. ASE impact on speed

This section summarizes outcomes based on studies carried out in Europe where the impact of ASE systems has been evaluated in detail. A number of studies have been conducted, which evaluate the impact of ASE on speed and crash rates. However, this paper dwells on the impact of average speed enforcement on speeding patterns. Soole et al. [7] compiled a concise literature survey of ASE evaluation in Europe. The aim of their research was to monitor compliance with posted speed limits on ASE sites. They also investigate the evidence of the effectiveness of ASEs through comparison with other countermeasures, driver perception and cost-benefit analyses. Previous studies in some enforcement sites reveal that average speed enforcement reduced the mean and 85th percentile speeds by up to 33%. In addition, speed variation from the posted speed limit was reduced with speeds typically below or at the posted speed limit [6][7]. Their findings support ASE as a complementary measure to existing speed management measures, which should focus on roads with historically high crash rates. Nevertheless, they conclude that ASE is a more reliable and cost-effective approach to speed enforcement, and is widely accepted by road users.

In the Netherlands, a study was conducted in 2005 on the A13 in Rotterdam with a posted speed limit of 80 km/h. Average speed in the enforcement site reduced from 100 km/h to 80 km/h. Reduction in speed variance and 85th percentile speed were also observed. Moreover offence rates dropped by 4% [11].

In Italy, an evaluation of all enforcement sites was conducted in 2009. Average speeds reduced by 16 km/h (corresponding to a 15% reduction) during the first year of operation. After the first year, average speeds further reduced by 9.1 km/h [7]. In 2011, a one week pre-installation and one week post-installation comparative study conducted on an 80 km/h road in Naples also showed positive impact. Average speed dropped by 9 km/h and speed variance dropped from 18.1 km/h to 12.1 km/h. In all cases, reductions were greater in free-flow conditions compared with peak periods [12].

In 2010, a series of ASE evaluations was conducted by speed check services on over 13 sites in England. Speed profiles three years before enforcement were compared with speed profiles three years during enforcement. The posted speed limits of the sites were between 30 mph and 50 mph. The 85th percentile speed dropped by an average of 14.4% for 11 sites, but increased for one site. Average speed reduced by an average of 12.5% for 10 sites, increased for two sites and remained unchanged for one site. The proportion of vehicles travelling above the speed limit reduced by an average of 30% [13].

According to the Western Cape government in South Africa, ASOD systems also have a positive effect on speeding patterns [5]. A macroscopic evaluation of the ASOD system on the R61 was conducted in 2012. Prior to ASOD enforcement, a total of 509 crashes were reported, 75 of which were fatal crashes. The specific time frame before ASOD implementation during which these crashes occurred is not mentioned. During ASOD enforcement, between November 2011 and November 2012, no fatal crashes were reported. The proportion of vehicles driving above the speed limit of 120 km/h dropped from 39% to 26%, and the percentage of vehicles driving below the speed limit increased from 61% to 74%. Moreover, the number of speed fines issued decreased from 2558 in January 2012 to 157 in August 2012 [5].

Previous studies focus on enforcement sites with little reference to the impact on control sites just outside enforcement sites. In addition, the impact of ASE systems is generalized for different modes of transport and vehicle types. In this paper, time differentiation (the period before and during ASOD enforcement) on enforcement and control sites is examined. Spatial differentiation (the impact on ASOD enforcement sites versus control sites) is also examined for a specific time frame. The analysis is carried out on passenger vehicles and minibus taxis based on data availability before and during ASOD enforcement.

III. METHODS

This section presents data sources, and discusses the methods applied to evaluate the extent to which ASOD systems have affected speeding patterns of passenger vehicles and minibus taxis on the R61.

A. Data sources

In order to conduct the evaluation effectively, commission dates and the precise location of cameras on roads must be known. This data was made available by relevant stakeholders. For this study, data was captured using tracking devices equipped with GPS receivers and cellular connectivity. Two independent data sets were used in this study. The paragraphs that follow discuss the data sources, accuracy and sample sizes.

The first data set was obtained from tracking devices installed in nine minibus taxis by Mix Telematics. Taxi owners were incentivised to have fleet monitoring devices permanently installed in their taxis. These devices provide GPS time, location and speed information at a frequency of 1Hz. This data set contains a total of 402 complete trips through the ASOD system between November 2013 and May 2014. Although the data was gathered from taxis within the same association, each

taxi owner has a set of contracted drivers who usually work across different minibus taxi associations. As such, the sample is a representative set of drivers who take the long distance route along the R61. Every weekend, minibus taxis from Cape Town travel to the Eastern Cape along this route. On a Friday night, about 300 minibus taxis use this route. The reader is referred to [8] for details on how the minibus taxi industry operates in South Africa.

The second data set was obtained from a database of historical tracking information from all vehicles tracked on TomTom Traffic Stats. Information was obtained from tracking devices, TomTom navigation devices, TomTom fleet management, and other TomTom solutions. The set contained over 2300 samples for the segments analysed before and during ASOD enforcement. This data was collected from a range of high quality data sources such as live PNDs (Personal Navigation Devices), in-dash navigation and business solutions, after which sophisticated data fusion was applied to achieve high accuracy and detailed road coverage.

B. Method of evaluation

The aim of this study is to investigate the impact of ASOD systems on speeding and compliance with speed limits inside and outside the ASOD enforcement site. Since the literature already provides evidence of crash reduction on the R61, the goal of this paper is to complement this evidence with detailed speed profile results. To achieve this, one enforcement site (with ASOD) and one adjacent control site (without ASOD) is evaluated. The control site was chosen such that its geometric characteristics and traffic conditions would be similar to those of the enforcement site. Figure 1 shows camera locations, the ASOD enforcement site and the control site for this study, and Table I is a summary of time frames and road characteristics. The date ranges were chosen while taking the implementation date of November 2011 and data availability into account.



Fig. 1: R61 evaluation section

Source: TomTom and Google earth view

Time differentiation was performed on the enforcement and control sites. This involved a ‘before’ and ‘during’ ASOD enforcement analysis on both the enforcement and control sites. Results from time differentiation on the enforcement site were expected to show reduction in travel speeds during ASOD enforcement. Similar results were also expected of the control site considering its proximity to the enforcement site.

Spatial differentiation was also performed with the aim of finding out the impact of ASODs on the control site

TABLE I: Summary of date ranges and road characteristics

Enforcement site	Passenger Vehicle		Minibus Taxi
	Before ASOD	Jan 2008-Jun 2011	-
With ASOD	Nov 2011-Dec 2013	Nov 2013-May 2014	
Distance (km)	51	71.6	
Speed limit (km/h)	120	100	
Control site	Before ASOD	Jan 2008-Jun 2011	-
	With ASOD	Nov 2011-Dec 2013	Nov 2013-May 2014
Distance (km)	52	54	
Speed limit (km/h)	120	100	

relative to the enforcement site. This involved ‘in’ and ‘out’ of ASOD analysis before and during ASOD enforcement. Results from spatial differentiation before ASOD enforcement were expected to be similar while results during enforcement were expected to be slightly different.

A further investigation was conducted for minibus taxis, which used the probe data to detect whether vehicles violated the system. Using camera locations, complete trips along the enforcement site were identified and analysed. Average speed along the enforcement site was calculated using distance travelled and travel time. The integrity of the analysis was ensured by excluding any trips with no probe data within one kilometre from both camera locations.

Eight months of probe data between November 2013 and June 2014 were available for minibus taxis along the R61. This means there was no sample data for minibus taxis before ASOD implementation in November 2011. Due to this data availability constraint, only spatial differentiation during ASOD enforcement has been performed for the taxis.

IV. RESULTS

In total, more than 2300 samples were used in the analysis of passenger vehicles and 402 trips were analysed for minibus taxis. Figure 2 shows the speed percentile plots for passenger vehicles and minibus taxis for each site and time. The figure shows an overall reduction in speed for passenger vehicles in both the enforcement and control sites during ASOD enforcement. There are, however, no significant differences between the speed profiles in enforcement and control sites for minibus taxis. Speed statistics obtained for passenger vehicles are summarized in Table II for time and spatial differentiation, while Table III shows the spatial differentiation results of taxis against those of passenger vehicles.

A. Passenger vehicles

As shown in Table II, the ASOD system appears to have had an impact on the speed profiles of passenger vehicles on the enforcement site. ASOD enforcement led to a reduction in mean speed by 5.5 km/h from 110.7 km/h before enforcement. The 85th percentile speed also reduced by 5 km/h, which corresponds to a 4% reduction. In addition, the percentile at the 120 km/h legal speed limit increased from 66% to 75%. This implies that passenger vehicles spent more time driving below the legal speed limit on the enforcement site.

The control site also showed positive results for passenger vehicles. Mean speed reduced by 6.9 km/h from 108.5 km/h before ASOD implementation. A 13 km/h reduction in the 85th

TABLE II: Spatio-temporal comparison for passenger vehicles

		N	Mean	V ₈₅	% ₁₂₀	% ₁₀₀	Δ_{mean}	Δ_{85}	Δ_{120}
Enforcement	Before ASOD	306	110.7	129	66	20			
	With ASOD	1389	105.2	124	75	30	-5.5	-5 (4%)	9
Control	Before ASOD	101	108.5	136	64	20			
	With ASOD	528	101.6	123	80	28	-6.9	-13 (10%)	16

Note: N = average sample size; Mean = mean speed; V₈₅ = 85th percentile speed; %₁₂₀ = 120 km/h percentile crossing; Δ = difference between During and Before. All speeds are in km/h.

TABLE III: Spatial differentiation for taxis versus passenger vehicles

		N	Mean	V ₈₅	% ₁₂₀	% ₁₀₀	Δ_{mean}	Δ_{85}	Δ_{120}	Δ_{100}
With ASOD	Enforcement	1389	105.2	124	75	30				
	Control	528	101.6	123	80	28	-3.6	-1	5	-2
With ASOD (Taxis)	Enforcement	402	110	128	60	14				
	Control	402	112	129	60	15	2.0	1	0	1

Note: N = number of trips for taxis and average sample size for passenger vehicles; Mean = mean speed; V₈₅ = 85th percentile speed; %₁₂₀ = 120 km/h percentile crossing; %₁₀₀ = 100 km/h percentile crossing; Δ = difference between Control and Enforcement. All speeds are in km/h.

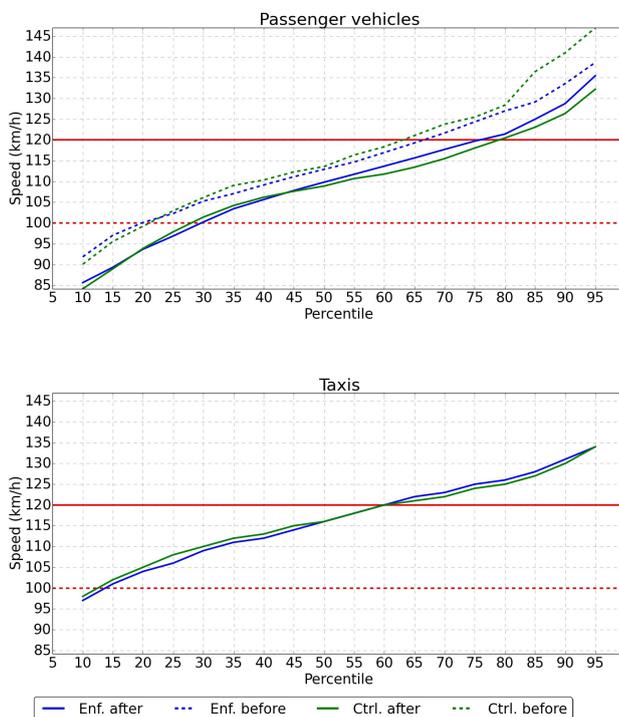


Fig. 2: Speed percentiles for passenger vehicles (top) and taxis (bottom)

percentile speed was also observed, corresponding to a 10% reduction. In addition, the percentile at the 120 km/h legal speed limit increased by 16% indicating that drivers adhered to the legal speed limit more often.

Time differentiation results for passenger vehicles showed that speed profiles on both the enforcement and control sites improved. However, the control site, which is some 20 km from the enforcement site, was characterized by a greater adherence to speed limits than the enforcement site. Given that other factors besides the ASOD system such as prolonged

road works may result in speed profile improvements, the time differentiation results need to be validated. This will be accomplished by examining the spatial differentiation results before ASOD enforcement.

Before ASOD enforcement along this road segment, speed profiles on the enforcement and control sites were quite similar. This is evident from the plots in Figure 2 and in Table II where a difference in mean speed of only 2.2 km/h is observed. In addition, percentiles at the 120 km/h speed limit differ by only two percentage points. These observations indicate that before ASOD implementation, driver behaviour in the enforcement and control sites was not only similar, but was also characterized by higher speeds. Hence, the spatial differentiation results before ASOD implementation show that speeding patterns were very similar, which indicates that the speed reduction observed in time differentiation can only be attributed to the ASOD system, in agreement with findings in [5]. In addition, while time differentiation analysis for passenger vehicles shows that the ASOD system had an effect on the enforcement and control sites, spatial differentiation analysis shows that it had a greater impact on the control site with a mean speed 3.6 km/h lower than that of the enforcement site. This suggests that the ASOD system along the R61 also influences drivers to comply with legal speed limits on the control site with no ASOD enforcement.

B. Minibus taxis

Currently, the posted speed limit for minibus taxis on the enforcement and control sites is 100 km/h. But looking at the speed percentiles in Figure 2, only about 15% of all recorded taxi speeds are within this legal speed limit. Further more, besides a lower variation in speed for the minibus taxis, the speed profiles of minibus taxis are very similar to, or higher than those of passenger vehicles. This finding conforms with studies in [8], which presents similar results for three other road sections. Also, Table III shows that the average speed of minibus taxis during ASOD enforcement in both sites is at least 110 km/h, which is very similar to the average speeds of passenger vehicles before ASOD implementation. Furthermore, it can be observed from the spatial differentiation

results that for minibus taxis, there appear to be no significant differences in driving behaviour at the enforcement and control sites due to very low percentile and mean speed changes. From these results, it is therefore evident that unlike passenger vehicles, minibus taxis are not influenced by the presence of the ASOD system along the R61. Also, the similarity between minibus taxi speeds during ASOD enforcement and passenger vehicle speeds before ASOD enforcement is an indication that performing time differentiation analysis on minibus taxis will show little or no significant changes

C. Further investigation on minibus taxis

An investigation on individual trips and speed distributions for minibus taxis shows that most drivers did not conform to the 100 km/h legal speed limit, which contradicts findings in [5]. Table IV shows a summary of system violations detected from probe data, for each tracked taxi. Results are expressed as a percentage of trips with an average speed beyond a certain threshold. The thresholds start at the 100 km/h speed limit and end at 120 km/h, with 5 km/h increments.

TABLE IV: Trip-based violations summary for taxis

Taxi	N	SL (%)	SL+5 (%)	SL+10 (%)	SL+15 (%)	SL+20 (%)
6000	74	81.1	71.6	62.2	52.7	32.4
6001	49	77.5	67.3	53	34.7	16.3
7000	32	75.0	56.2	31.2	15.6	0.0
7001	53	90.5	75.5	56.6	30.2	13.2
3001	56	80.3	76.8	64.3	50.0	21.4
1000	60	83.3	75.0	58.3	35.0	16.6
5000	30	83.3	76.6	56.6	46.6	33.3
4000	28	85.7	78.6	67.8	35.7	10.7
1001	20	70.0	60.0	35.0	20.0	15.0

Note: N = number of complete trips through ASOD system. SL = Speed limit of 100 km/h. $SL+10$ = 110 km/h. $SL+10$ (%) is the percentage of trips with average speed greater than 110 km/h.

The results show that at least 70% of trips taken by each taxi violate their 100 km/h legal speed limit, and for some taxis, close to 34% of their trips violate the 120 km/h speed limit for passenger vehicles. While these results show that ASOD enforcement has little or no impact on minibus taxis, they also support findings in [14] on the impracticality and enforcement difficulties associated with differentiated speed limits. Furthermore, interviews with some of the drivers revealed that although they know the 100 km/h speed limit, they nevertheless consider 120 km/h as the limit which governs their choice of speed.

V. CONCLUSION

Even in the South African context, the effectiveness of ASOD systems in speed reduction and compliance with speed limits is unquestionable, especially for passenger vehicles. Although the average speed of passenger vehicles along the R61 was lower than their 120 km/h speed limit before ASOD implementation, a significant reduction in speed was still observed during ASOD enforcement. In addition, results for passenger vehicles show that the ASOD system also improved speed on the control site without ASOD. On the other hand, speeds of minibus taxis were not similarly reduced. Speed profiles of minibus taxis during ASOD enforcement on the enforcement and control sites were very similar, indicating that the system does not affect the taxis. Since many taxis

have been identified for speed-related violations by this ASOD system in previous months, this discrepancy is peculiar. In addition, average speeds measured from probe data also show that most trips violate the system. The reason for such persistent violation by minibus taxis could either be due to discriminatory law enforcement which lets taxis drive above their legal speed limit, or due to failure of the ANPR cameras in detecting the taxis altogether. Investigation on these reasons were beyond the scope of this paper, and as a result, reserved for future work. Further research will also investigate the impact of ASOD systems on other South Africa roads for the same vehicle types considered in this study. This will also include survey responses from the drivers, examining their knowledge on how the ASOD system operates.

VI. ACKNOWLEDGEMENT

The authors would like to acknowledge MTN, Mix Telematics and TomTom for their financial and technical support.

REFERENCES

- [1] E.D. Richter, T. Berman, L. Friedman, G. Ben-David, "Speed, road injury and public health," *Annual review of public health* 27 (1), 125-152, 2006.
- [2] R. Elvik, A. Høy, T. Vaa, M. Sørensen, "The handbook of road safety," *Emerald Group Publishing Limited*, 2nd ed., Bingley, UK, 2009.
- [3] L. Aarts, I. van Schagen, "Driving speed and the risk of road crashes: a review," *Accident Analysis and Prevention* 38 (2), 215-224, 2006.
- [4] Road Traffic Management Corporation, "Road Traffic report," *Department of Transport: Republic of South Africa*, March 2011.
- [5] Launch of Average Speed Enforcement Technology, (Western Cape Government: safely home), [online] 2011, Available: <http://safelyhome.westerncape.gov.za/campaigns/818> (Accessed: 8th August 2014).
- [6] D. Soole, J. Fleiter, B. Watson, "Point-to-point speed enforcement: Recommendations for better practice," *Australasian Road Safety Research, Policing and Education Conference*, Brisbane, Queensland, 2013.
- [7] D. Soole, J. Fleiter, B. Watson, "Effects of average speed enforcement on speed compliance and crashes: A review of the literature," *Accident Analysis and Prevention* 54, 46-56, 2013.
- [8] M.J. Booysen, N.A. Ebot Eno Akpa, "Minibus driving behaviour on the Cape Town to Mthatha route," *Southern African Transport Conference*, 2014.
- [9] P. Liu, X. Zhang, W. Wang, C. Xu, "Driver response to automated speed enforcement on rural highways in China," *Journal of the Transportation Research Board of the National Academies*, Washington D.C., No. 2265, 109-117, 2011.
- [10] F.K. Afukaar, "Speed control in developing countries: issues, challenges and opportunities in reducing road traffic injuries," *Injury Control and Safety Promotion*, 10:1-2, 77-81, 2003.
- [11] C. Stefan, "Automatic Speed Enforcement on the A13 Motorway (NL): Rosebud WP4 Case B Report," *Austrian Road Safety Board (KfV)*, Austria, 2005.
- [12] E. Cascetta, V. Punzo, "Impact on vehicle speeds and pollutant emissions of an automated section speed enforcement system on the Naples urban motorway," *Paper Presented at the TRB 2011 Annual Meeting*, 2011.
- [13] Speed Check Services, Average Speed Enforcement Solutions: Safer, Smoother, Greener, Fairer. Speed Check Services, London, 2010.
- [14] C.J. Bester, "Differentiated speed limits that will work", *South African Transport Conference*, CSIR, Pretoria, July 2012.

Visibility Improvements through Information Provision Regarding Sun Glare:

A Case Study in Cape Town

Marianne Vanderschuren (*Centre for Transport Studies, University of Cape Town*)
Nothando Khumalo (*Centre for Transport Studies, University of Cape Town Affiliation*)
Rondebosch, South Africa
Marianne.Vanderschuren@uct.ac.za

Abstract—This paper describes a GIS-based methodology to determine direct sunlight exposure for the road network in Cape Town. Thus, for an arbitrary position in the roadway alignment of the study area, the amount of risk for drivers resulting from vision impairment as consequence of sun glare was analysed for the equinox and solstice days. The results of this procedure are hillshade and road segment maps of the four days illustrating the areas and streets where direct sunlight exposure is experienced. A case study carried out in one of the Cape Town streets proved that the methodology produces valid and reliable results. Therefore, results from this procedure can be an informative dimension to consider when evaluating existing roads or different layout and alignment alternatives for new roads. Moreover, the methodology can be incorporated into car navigation systems to provide automated real time sun glare risk information to drivers.

Keywords— *Direct Sunlight Exposure; ArcGIS; Cape Town*

I. INTRODUCTION

There are numerous potential explanations of traffic crashes and it is not surprising that so many dimensions appear important, since driving is such a complex task (Plainis et al., 2006). Visibility conditions have been identified by various authors as being an important environmental factor. In most cases, visibility refers to night and rainy conditions, both influencing the severity and the rate of crashes (Clarke et al., 2006; Konstantopoulos et al., 2010). Sun glare is mentioned in the literature as a visibility factor (Pande and Abdel-Aty, 2009; Mitra and Washington, 2012; Dozzaa and Paneda González, 2013). Staubach (2009) evaluated factors correlated with traffic accidents as a basis for evaluating Advanced Driver Assistance Systems (ADAS). Sun glare was identified as a factor correlated with causation of crossroad accidents for the driver, in over half of all cases (52%).

According to the National Highway Traffic Safety Administration (NHTSA) of the U.S. Department of Transportation, sun glare is the official cause of a fraction of fatal crashes across the country, which is 195 in 56 793 (0.34%) (Hastings, 2012). Data compiled by the Abu Dhabi Traffic Department showed that sun glare was blamed for 22 minor crashes in the capital during the first eight months of the year 2010 (Salama, 2010). Furthermore, a number on accident cases were reported in the USA where sun glare was

identified as the cause of the accident. One of these occurred in Colorado Springs, where four Coronado High School students landed in hospital in the month of September after they were hit by a van while crossing the street on the cross walk (Stone, 2007). The driver, who was going into an easterly direction into the sun, claimed to have been blinded by the sun. In a different case, a truck driver in Syracuse New York struck a female pedestrian while she was crossing the streets, causing her death (Ha, 2011).

Research by Hagita and Mori (2013) indicates that the angle of the sun is most potent to drivers during the times when the sun is closest to the horizon, which is dawn and dusk; typically between 07h30-09h30 and 17h00-18h30, respectively. In agreement, Mitra (2008) also disclosed that the traffic accident rates at dusk and dawn are higher than the rates at other times. This was concluded in an analysis of traffic accident data, sunset and sunrise time, and road travel directions in Arizona, America. Several investigations have shown that high traffic volume is linked with an increase of collision rate and a decrease of fatality number (Auffray, 2007; Abdel-Aty and Radwan, 2000). As such, it is highly likely that roads with low traffic volumes have considerably less sun glare risk, compare to roads that carry peak traffic at volumes that produce higher speeds. Although this is debatable considering that the vehicle speeds in a congested road are lower compared to those of a free flowing road. In slower moving traffic, drivers experiencing sudden blinding, as a result of sun glare, have a relatively longer reaction time and a lower crash/impact severity as a result of the low speeds, compared to free-flowing traffic; meaning the risk is greater in the latter. Glare resulting from direct sunlight exposure can be painful to the eye of the observer and potentially very distracting to the driver in terms of visibility (Auffray, 2007). Consequently, this distraction to the driver has well under-stood adverse effects, not only on the safety of the driver but on adjacent drivers as well.

The key objective of this study was to develop a method that determines, for an arbitrary position on a roadway alignment, what times of the year/day drivers could be faced with the risk of vision impairment as a result of sun glare. The study aimed to present a methodology which can model direct sunlight exposure for the Cape Town road network with the use of ArcGIS. Based on the assumptions and limitations, the outcome of this methodology gives an overview of the

vulnerability of road network segments to accidents as a result of sun glare.

II. METHODOLOGY

A number of factors influence the occurrence of sun glare conditions. In this study, three factors were identified as key influences to sun glare occurrence: road network (geometric design), physical environment (topography and terrain profile), and sun position (azimuth and altitude). The wrong combination of these factors could result in hazardous conditions due to increased sun glare risk, which is actually, to a larger extent, a road safety risk. Taking into account the need for a tool with the ability to combine spatial data and sun position data, ArcGIS (Geographical Information Systems (GIS) technology) was considered the most suitable tool for this study. Simple modules for computing sunlight exposure, which require only the surface data and the sun position angles, are available in GIS programs. Such modules include the hillshade tool in ArcGIS, which was utilised here. The tool identified road segments exposed to direct sunlight in the Cape Town road network.

A. Research Tool

ArcGIS is Geographic Information System software, produced by ESRI, which allows people to collect, organise, manage, analyse, communicate, and distribute geographic information (www.esri.com). The role of ArcGIS in applications such as urban planning and transportation provides users, managers and decision makers, powerful tools for solving complex spatial problems.

Hillshade is an ArcGIS tool that was employed to model the effects of solar exposure. Hillshade is a shaded relief technique where a lighting effect is added to a map based on elevation variations within the landscape (<http://landtrustgis.org/>). The hillshade function intends to mimic the sun's effects – illumination, shading and shadows – on hills and canyons, thus, obtaining hypothetical illumination of a surface by determining illumination values for each cell in a raster (Hegazy and Effat, 2011). This function uses the latitude and azimuth properties to specify the sun's position, which are the function's inputs including a Digital Elevation Model (DEM) and a z-factor. A DEM is the presentation of continuous elevation values over topographic surface by a regular array of z-values, referenced to a common datum, DEMs are typically used to represent terrain relief (<http://support.esri.com/>). A hillshade model is a derivative of a DEM that stimulates relative solar insolation for each grid cell based on its slope, aspect and the position of the sun (as defined by elevation and azimuth angle) (Bricher et al., 2008). The azimuth is the sun's relative positions along the horizon, and is expressed in positive degrees ranging from 0 to 360, measured clockwise from north. The altitude is the sun's angle of elevation above the horizon, and is expressed in positive degrees ranging from 0 to 90° - with 0° at the horizon and 90° directly overhead (<http://www.esri.com/>).

B. Case Study

The chosen study area for this research is the City of Cape Town (CoCT), in South Africa, which occupies the south-western most point of Africa. Cape Town is a legislative capital of South Africa and capital of the Western Cape Province, and is located at 33°55'31"S latitude and 18°25'26"E longitude in the Southern Hemisphere. The city is the second largest city in South Africa based on population, and is the largest in land area at 2 455km² (<http://www.britannica.com/>). The City centre lies embedded between Table Mountain, Devils Peak, Lions Head and Signal Hill on the one side and borders on the Table Bay and the Atlantic Ocean on the other side.

C. The Data

The data used in this study was categorised into spatial data (DEM, road network and topography), sun position data (azimuth and altitude) and sun cone data. The study area DEM, obtained from the University of Cape Town GIS laboratory, was derived from 10m-interval (spatial resolution) contour lines. The slope and aspect of the DEM played a key role in the creation of hillshade layers. The road network layer was derived from the topographical map.

A study of driver impairment situations as a result of sun glare, Jurado-Pina and Pardillo-Mayora (2009) identified two values of the angle of glare to characterise problem situations, i.e. 19° and 25° (altitude). After some tests, this study adopted the same values for the altitude sun cone.

The analysis instruments provided by the ArcGIS Spatial Analyst hillshade tool allow a map drawing and an analysis of the sun's effects on a geographical area during a certain specific time frame. Two astronomical events (solstice and equinox) each of which occur twice a year, and have the added benefit of seasonal variation, were selected for use in this research. The Autumnal Equinox (AE) and Spring Equinox (SE) occur on the 20th of March and 22nd of September (2014) in the Southern Hemisphere, respectively. While the Winter Solstice (WS) and Summer Solstice (SS) occur on the 22nd of June and 22nd of December.

Subsequently, Sun position data (azimuth and altitude) for these four days, for both the morning (AM) and afternoon (PM) snapshots, was obtained from the Astronomical Applications Department of the U.S. Naval Observatory server (aa.usno.navy.mil/). A decision was made to select time snapshots – for the four days – whose altitude values are defined by the vertical sun cone (19° and 25°). This means each scenario/time snapshot modelled in this study either had a 19° or 25° altitude value.

Overall, a total of 16 scenarios/time snapshots were formulated for this model. First and foremost, there are the four days which are split into 2 equinox days and 2 solstice days. For each of these four days, there is the morning (AM) and afternoon (PM) scenario. Each of the AM and PM scenarios are further split into two time snapshots each of which is for the 19° and 25° altitudes. Table I gives an example of the sun position data for equinox days, which has a total of 8 scenarios (time snapshots).

TABLE I. EQUINOX SUN POSITION

Date (2014)	Day	Time Period	Time	Altitude (°)	Azimuth (°)
20 Mar	AE	AM	08h30	19.9	76.2
			08h55	24.8	72.1
		PM	16h55	24.4	287.7
			17h20	19.4	283.6
23 Sept	SE	AM	08h15	19.8	76.1
			08h40	24.8	72.0
		PM	16h40	24.4	287.5
			17h05	19.5	283.4

III. GENERAL FINDINGS

The Cape Town road network was separated into small straight line road segments with constant geometry. Several statistical tests were applied while analysing the results from the model. The data analysis and findings presented entails a discussion of the percentage of road segments during each day, as well as a discussion of the statistical analysis outcome.

In comparing the percentage values of road segments exposed to direct sunlight in the morning and afternoon period, the apparent trend is a decrease from the morning to the afternoon period in the AE and SE scenarios, and an increase for the same change in the WS and SS scenarios. For example, at 25° altitude, the percentage of road segments in the autumnal equinox changes from 21% in the morning to 15% in the afternoon. At the same altitude the percentage of roads at risk in the summer solstice increases from 16% in the morning period to 21% in the afternoon period. Presumably this implies that the solar illumination in the afternoon period of the winter and summer solstice has a relatively greater effect than the morning period, which means the opposite applies to the autumnal and spring equinox. The SE boasts the same percentage values as the AE, for both AM and PM. With 13% and 15% for the AM and PM period, it is not surprising to learn that the winter solstice has the least amount of road segments exposed to direct sunlight.

A statistical analysis comparing the morning and afternoon period was carried out using the Wilcoxon signed rank test as it is a non-parametric equivalent of the parametric t-test, which cannot be used if the data are not normally distributed, as is the case. The analysis produced a p-value of 0.0117 for the autumnal and spring equinox, implying there is a significant difference in the median amount of the average sunlight between AM and PM. On the contrary, the WS data produced a p-value of 0.5779 which is greater than the 5% statistical significance level, implying there is no evidence for a statistically significant difference in the median amount of average sunlight between AM and PM. The significance level is a boundary at which to assume there is evidence to reject the null hypothesis, i.e. if $p \leq 5\%$. On the other hand, if $p > 5\%$, it implies there is insufficient evidence of a difference in the results and the null hypothesis is accepted, which is the case with the WS. The p-value for the SS is 0.0221. Similar to the AE and SE, it also implies a significant difference in the median amount of average sunlight between AM and PM. However, the AE and SE have relatively stronger evidence

considering that the evidence against the null hypothesis in favour of the alternative increases with a decreasing p-value. An overall comparison of all the four days and times, using a linear mixed-effects model, revealed a number of findings. The streets were modelled as random effects as there are multiple measurements per street. AM and PM were also interacted with day to see if there is an effect. The output revealed that, on average there is less sunlight in the afternoon compared to the morning, as the effect of PM is negative and statistically significant (see Table II).

TABLE II. LINEAR MIXED EFFECT MODEL RESULTS

Percentage	Effect Size	Std. Err.	p-value
SE	-0.16	1.86	0.933
SS	-11.15	2.05	<0.001
WS	-10.10	2.22	<0.001
PM	-11.32	2.05	<0.001
SE:PM	0.16	2.72	0.954
SS:PM	22.32	3.06	<0.001
WS:PM	11.75	3.16	<0.001
Constant	53.90	3.04	<0.001
Between Street	0.32		
Within-street/error	0.17		

IV. MODEL VALIDATION

On the 15th of August, 2014, a video footage of four Freeway and Primary arterial roads in the Cape Metropolitan CBD were taped for the morning and afternoon periods during the time periods when sun was considered to be at its worst – which in this case was defined by the 19°/25° sun cone. The investigation was carried out using a Road Eye JS-300 camera, which is an intelligent electronic device designed to record driving parameters. Footage was recorded while travelling at a speed approximately between 40 and 60km/h. The terrain around these four roads is relatively flat.

In the morning the 19°/25° sun cone occurred between 09h15 and 09h45, while in the afternoon it occurred between 15h55 and 16h30. Considering that these times were well within an hour or so after sunrise (07h24) and before sunset (18h16), a decision was made to record sun glare footage between these periods. The clear weather condition on this day also made for a suitable recording environment. **Figure 1** provides an example of the footage collected. While the top image clearly indicates a sun glare challenge, the bottom image indicates that high buildings can influence sun glare exposure in practice.



Fig. 1. Buitenkant Street Snapshot (NE bound)

Using the ArcGIS Hillshade tool for the same day, time and area, the calculations provided the same result, i.e. sun glare exposure in Buitenkant Street (see Figure 2). According to the map in Figure 2, the street segments affected by direct sunlight exposure at this time of day are northeast-orientated, as indicated by the light-blue coloured streets. For instance, northeast-bound traffic in the Buitenkant segment between Roeland Street and Darling Street is exposed to direct sunlight.

The current data for this area does not include the buildings. However, even if exposure to sun glare might be reduced at times, the overall risk is still prevalent. The application of a road user warning system, through in vehicle or street based ITS systems, would reduce the road safety risk identified.



Fig. 2. Morning Period Segment Distribution Map for Buitenkant Street

V. CONCLUSIONS

Sun glare can be a nuisance and increase the road safety risk. The position of the sun and the angle of the rays may render sun visors useless. Accordingly, terrain analysis using DEM-based hill shading has led to the discovery of a method to identify roads vulnerable to sun glare conditions. The identification of road segments at risk is useful information with regards to the implementation of mitigation measures, such as visual barriers or warning signs, to prevent accidents.

The implementation of this method to the Cape Metropolitan road network showed that the AE and SE both have approximately 14.7% of road network exposed to direct sunlight. The WS and SS, on the other hand, have approximately 12.2% and 15.2% of the road network at sun glare risk, respectively. With the exception of the WS, statistical analysis of the results revealed that there is a significant difference in the amount of sunlight exposure between the morning and afternoon time periods in the other three days (AE, SE and SS). According to these results, on average there is less sunlight in the afternoon compared to the morning. Looking at the individual days, the AE and SE are in accord with this outcome – they have less sunlight in the afternoon. However, the WS and SE actually have more sunlight exposure in the afternoon than the morning. Overall, the AE and SE have the most sunlight exposure, while the WS has the least.

The main assumption of the study was whether increased exposure to sun glare entails an increased risk of accidents.

Various authors have proven the relationship. Another assumption in the study was the identification of a sun cone ranging between 19° and 25° altitude, which according to Jurado-Piña and Pardillo-Mayora (2009) characterise vision impairment. The validation has proven that this assumption was valid. Furthermore, a horizontal sun cone threshold ($\pm 15^\circ$) was adopted for azimuth and slope filtering.

The model does not allow for atmospheric effects (cloud, rainfall etc.) and the presence of buildings or trees affecting the direct sunlight. All the assumptions and omissions influence the results and provide opportunities to extent the study. On a different note, with respect to future research, a 24-hour calculation of sunlight exposure can be carried out, whereby the duration of direct sunlight exposure per “spot” can be used to evaluate the sun glare risk.

Outcomes from the proposed methodology can be applied to both existing and new road designs. With respect to the design process for new roads, it would be an informative dimension to consider particularly when evaluating different layout and alignment alternatives. Aside from the use of mitigation measures, such as visual barriers (e.g. row of trees), Intelligent Transport Systems (ITS) could be employed. Considering that the model produces results for a specific time, automated real time warning systems could be implemented for the identification of high risk conditions and road segments in terms of direct sunlight exposure. In a similar study, Chalkias et al. (2013) proposed incorporating a methodology like this in car navigation systems, in order to provide additional real-time information to drivers. The incorporation of GIS technology with Global Positioning Systems could enable the development of optimal automated real time warning systems. All in all, ITS research could potentially provide endless opportunities for the mitigation of sun glare conditions in line with this methodology.

REFERENCES

- [1] M.A. Abdel-Aty, A.E. Radwan, Modeling traffic accident occurrence and involvement. In *Accident Analysis and Prevention*, 32(5), pp.633-642, 2000.
- [2] ArcGIS for Desktop, Available from: <<http://www.esri.com/>>. [13 March 2014].
- [3] B. Auffray, Effect of the Sun Glare on Traffic Flow Quality, Terwilliger, Portland, Oregon, USA, 2007.
- [4] P.K. Bricher, A. Lucieer and E.J. Woehler, Population trends of Adélie penguin (*Pygoscelis adeliae*) breeding colonies: a spatial analysis of snow accumulation and human activities, *Polar Biol*, 2008, Vol. 31, pp. 1397–1407.
- [5] Cape Town, Encyclopaedia Britannica. Encyclopaedia Britannica Online Academic Edition, Encyclopædia Britannica Inc., 2014. <<http://www.britannica.com/EBchecked/topic/93686/Cape-Town>>. [25 May 2014].
- [6] C. Chalkias, A. Faka and K. Kalogeropoulos, Assessment of Direct Sun-Light on Rural Road Network through Solar Radiation Analysis Using GIS. *Open Journal of Applied Sciences, Scientific Research*, 2013. DOI:10.4236/ojapps.2013.32030.
- [7] D.D. Clarke, P. Ward, C. Bartle, and W. Truman, Young driver accidents in the UK: the influence of age, experience, and time of day, *Accident Analysis and Prevention*, 2006, Vol. 38 (5), pp. 871–878.
- [8] M. Dozzaa, and N. Pa'neda González, Recognising safety critical events: Can automatic video processing improve naturalistic data analyses? *Accident Analysis and Prevention*, 2013, Vol. 60, pp. 298–304.
- [9] Y. Ha, New York Court Finds Driver Blinded by Sun Glare May be Liable in Accident, *East News, Insurance Journal*, 2011. Available from: <<http://www.insurancejournal.com/news/east/>>. [13 August 2014].
- [10] K. Hagita, and K. Mori, The Effect of Sun Glare on Traffic Accidents in Chiba Prefecture, Japan, *Proceedings of the Eastern Asia Society for Transportation Studies*, Vol.9, 2013.
- [11] P. Hastings, As sun sets, risk of crashes rises, 2012. *The Columbian* 14 September. Available from: <<http://www.columbian.com/news/>>. [10 August 2014].
- [12] M.N. Hegazy and H. Effat, Exploring the Egyptian Terrain Characteristics from Space for Strategic Planning, Survival and Sustainability, *Environmental Earth Sciences*, Springer-Verlag Berlin Heidelberg. DOI 10.1007/978-3-540-95991-5_75, 2011.
- [13] R. Jurado-Piña, and J.M. Pardillo-Mayora, Methodology to Analyze Sun Glare Impact on Highway under Prolonged Exposure, *Journal of Transportation Engineering-Asce*, 2010, Vol. 136(12), pp. 1137-1144.
- [14] P. Konstantopoulos, P. Chapman and D. Crundall, Driver's visual attention as a function of driving experience and visibility. Using a driving simulator to explore drivers eye movements in day, night and rain driving, *Accident Analysis and Prevention*, 2010, Vol. 42, pp. 827–834.
- [15] S. Mitra, Investigating Impact of Sun Glare on Transportation Safety, TRB 87th Annual Meeting, Transportation Research Board, Washington, DC., 2008
- [16] S. Mitra, and S. Washington, On the significance of omitted variables in intersection crash modelling, *Accident Analysis and Prevention*, 2012, Vol. 49, pp. 439–448.
- [17] A. Pandea and M. Abdel-Aty, 2009. A novel approach for analyzing severe crash patterns on multilane highways, *Accident Analysis and Prevention*, 2009, Vol. 41, pp. 985–994.
- [18] S. Plainis, I.J. Murray, and I.G. Pallikaris, Road traffic casualties: understanding the night-time death toll, *Injury Prevention*, 2006, Vol. 12, Issue 2, pp. 125–138.
- [19] S. Salama, Blinding sun caused 22 car crashes in eight months. *The Gulf News* 19 October 2010. Available from: <<http://gulfnews.com/news/gulf/>>. [16 August 2014].
- [20] M. Staubach, Factors correlated with traffic accidents as a basis for evaluating Advanced Driver Assistance Systems, *Accident Analysis and Prevention*, 2009, Vol. 41, pp. 1025–1033.
- [21] M. Stone, The Sun's Glare and Car Accidents. *KKTV* 09 October 2007. Available from: <<http://www.kktv.com/home/headlines>>. [10 August 2014].

Recognition of driving manoeuvres using smartphone-based inertial and GPS measurement

Jarrett Engelbrecht, Marthinus J. Booysen, Gert-Jan van Rooyen
MTN Mobile Intelligence Lab
Department of Electrical and Electronic Engineering
Stellenbosch University
Stellenbosch, South Africa, 7600
Email: {15608301,mjbooyesen,gvrooyen}@sun.ac.za

Abstract—The ubiquitous presence of smartphones provides a new platform on which to implement sensor networks for ITS applications. In this paper we show how the embedded sensors and GPS of a smartphone can be used to recognize driving manoeuvres. Smartphone-based driving behaviour monitoring has applications in the insurance industry and for law enforcement. The proposed solution is suitable for real-time applications, such as driver assistance and safety systems. An endpoint detection algorithm is used on filtered accelerometer and gyroscope data to find the start- and endpoints of driving events. The relevant sensor data is compared against different sets of manoeuvre signal templates using the dynamic time warping (DTW) algorithm. A heuristic method is then used to classify a manoeuvre as normal or aggressive based on its speed and closest matching acceleration and rotation rate templates.

I. INTRODUCTION

Worldwide, more than a million deaths are caused by road accidents per year [1]. The World Health Organization predicts that road fatalities will rise to become the fifth leading cause of death by 2030 [1]. Research done in the United States shows that, in more than 50% of fatal road accidents, unsafe driving behaviours were involved [2]. Road accidents are caused by a variety of factors, but aggressive driving behaviour is one of the major causes.

In the last decade, various companies have been developing solutions to monitor a vehicle and its driver's behaviour [3]–[5]. However, these solutions are expensive and intended for fleet management, and there is little incentive for individuals to buy them. However, the increasingly ubiquitous presence of smartphones – with their variety of sensors – presents the possibility to easily implement vehicle monitoring systems on a large scale.

Most modern smartphones have a variety of embedded sensors — typically an accelerometer and gyroscope, and light, proximity and magnetic sensors, as well as a microphone, camera and Global Positioning System (GPS). This variety of sensors make many sensing applications possible. An example of such an application is gesture recognition, which is used to answer a call when bringing the phone to one's ear, or paging through a document by the wave of a hand [6], [7]. In a similar way, different activities such as walking, running, cycling and driving can be detected and classified using the inertial sensors of a phone that is carried in a user's pocket [8].

Vehicle monitoring is an attractive sensing application for smartphones. For instance, drivers can be monitored to make them aware of their potentially dangerous driving behaviour. Anonymous participatory sensing could also enable identifying areas where accidents are more likely to occur [9].

Smartphones' connectivity also allows for the implementation of other vehicle monitoring features, such as traffic monitoring, traffic re-routing and accident reporting. Accident detection is possible using only the sensors in a modern smartphone, as shown by White et al. [10]. The swift automatic reporting of road accidents to authorities can prevent fatalities by minimizing the response time of emergency services. Additionally, using a machine-to-machine (M2M) communication platform would allow the redirection of other drivers away from an accident. Notifying drivers that they are approaching an accident scene could also increase their alertness and warn them to slow down, thereby preventing further accidents.

The remainder of this paper is organized as follows: Section 2 presents the current state of the art of smartphone-based monitoring systems; Section 3 describes the design of a proposed driving manoeuvre recognition system; Section 4 provides the experimental approach and results; and Section 5 presents the concluding remarks.

II. STATE OF THE ART

In this section, a brief overview is given of the current literature on smartphone-based monitoring systems used in vehicles. The literature mentioned is mostly relevant to driver behaviour monitoring, and some systems also employing road condition monitoring. The techniques and sensors used in the more recent projects are listed in Table I.

Johnson and Trivedi [9] developed one of the first complete driver behaviour monitoring systems on a smartphone. Their system can detect and classify a number of aggressive and non-aggressive driving manoeuvres when placed in a vehicle, by only using the internal accelerometer, gyroscope, magnetometer and GPS of a smartphone. Although the system can identify aggressive driving manoeuvres, it does not draw any conclusions from them. Their intent is to use the system to support a holistic driver assistance system (DAS) by providing it with additional information.

TABLE I
SUMMARY OF TECHNIQUES AND SENSORS USED BY SMARTPHONE-BASED VEHICLE MONITORING SYSTEMS.

Reference	Detection technique	Sensors used
Mohan [11]	pattern matching, orientation calibration	accelerometer, microphone, GPS
Dai [12]	pattern matching, orientation calibration	accelerometer, gyroscope
Johnson [9]	endpoint detection, DTW	accelerometer, gyroscope, magnetometer, GPS
Eren [13]	endpoint detection, DTW, Bayesian classifier	accelerometer, gyroscope, magnetometer
Fazeen [14]	pattern recognition	accelerometer, GPS
White [10]	pattern matching	accelerometer, microphone, GPS

Eren et al. [13] also implemented a smartphone-based driving manoeuvre detection system similar to Johnson and Trivedi’s [9] approach. However, they expanded the system by adding a driving style characterization feature that labels a person’s driving style as either safe or unsafe with a given probability.

Dai et al. [12] developed a smartphone-based system that specifically detects drunk driving. This is achieved by detecting and positively identifying a combination of dangerous driving manoeuvres associated with drunk driving.

Fazeen et al. [14] implemented a DAS entirely on a smartphone. The system records and analyses various driver behaviours and external road conditions, and advises a driver on dangerous vehicle manoeuvres. In addition to the driver behaviour monitoring feature, Fazeen et al. [14] also added a road condition characterization and mapping feature to their system that uses a smartphone’s GPS and accelerometer.

Mohan et al. [11] developed a comprehensive road and traffic monitoring system, named Nericell, which also employs smartphone sensors to detect certain driving manoeuvres and road conditions.

White et al. [10] developed the WreckWatch accident detection system for a smartphone. It detects a vehicle collision by applying threshold filtering to accelerometer and microphone samples from the smartphone. Data recorded before and during an accident is sent via GSM to a centralized server. Important information about an accident can then be relayed to the relevant authorities from a stored database on the server.

The literature distinguishes between driving manoeuvre recognition and driving behaviour classification. A system could detect various manoeuvres, but not necessarily infer anything from them, whereas another system may be able to classify a driver’s behaviour from detected driving manoeuvres. These different systems demonstrate the variety of driving behaviour classifications that can be made. A person’s normal driving style can be classified as safe or risky, fuel-efficient or inefficient, skilled or unskilled — and recommendations can be given accordingly to improve their driving. On the other hand, a person’s driving behaviour can sometimes differ from normal due to certain circumstances. A person could be

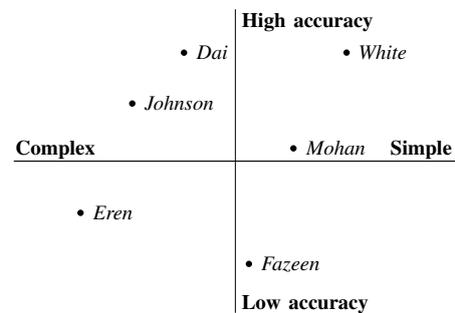


Fig. 1. Qualitative comparison of the systems’ accuracy versus simplicity.

driving under the influence of alcohol, drugs or other sensory impairments. In such situations drivers could be warned of their dangerous behaviour, and with their consent, the relevant authorities could be notified of their location.

A. Accuracy versus simplicity

It is impractical to quantitatively compare the performance and power consumption of the different systems. All of the systems were implemented on different smartphones that have varied sensors and computing power. The test studies were performed in various countries with different road and traffic conditions. Their methods of establishing the ground truth for tests were not necessarily the same and could vary due to subjectivity.

Figure 1 shows a qualitative comparison of the accuracy versus simplicity of the different systems. A system that achieves high detection accuracy with a simple algorithm is considered superior. The assumption is that a simpler system uses less resources and therefore consumes less power. The experimental and empirical test results of the systems as given in each paper were used to compare detection accuracy, although the testing procedures differed as mentioned. The perceived simplicity of each system is based on what each system is trying to detect, what sensors it uses and how its algorithms function.

WreckWatch of White et al. [10] is empirically proven to be virtually 100% accurate and is the simplest system, because it only detects accidents and nothing else. The road condition monitoring feature of Mohan et al. [11] is more accurate than that of Fazeen et al, and its implementation is simpler.

The drunk driving detection of Dai et al. [12] is the most accurate, achieving a false negative rate of virtually 0%. Dai et al. [12] implemented a simple yet effective pattern matching approach that requires very little computation. Essentially, only the difference in subsequent values on the relative longitudinal and latitudinal axes are calculated. If the difference exceeds a certain threshold, an aggressive driving manoeuvre is assumed. The algorithm used by Nericell of Mohan et al. [11] works in a similar manner. Both systems consume less than 12% of the phone’s battery life-cycle.

In contrast, Johnson and Trivedi [9], as well as Eren et al. [13], implemented a more complex pattern recognition approach derived from speech recognition techniques. Their

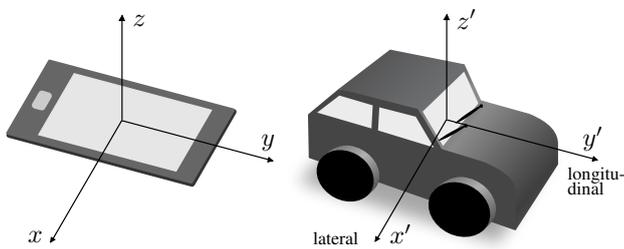


Fig. 2. Smartphone and vehicle coordinate system.

systems perform well, achieving a true positive rate of 97% and 93%, respectively. Although it can not be explicitly proven here, the simpler approaches are likely to consume less power while achieving similar performance to the more complex approaches. Arguably, Dai et al. [12] accomplished the same functionality as Eren et al. [13] with a simpler algorithm, as both systems can infer a certain aspect of a driver's behaviour from detected driving manoeuvres.

The systems of Mohan et al. [11] and Dai et al. [12] were developed on hardware and software that are now considered obsolete, yet their systems were simple and accurate. This suggests that in the last decade, the accuracy of embedded sensors used in smartphones has either not improved significantly, or it does not result in better manoeuvre recognition. The computing power and efficiency of modern smartphones, however, has increased dramatically, which provides headroom for more complex solutions. Therefore there is still merit in implementing a more complex approach as used by Johnson and Trivedi [9]. Especially if the accuracy could be improved to such an extent as to have very few false negatives (FN) or false positives (FP).

B. Contributions and best practices

In terms of contributions made, Dai et al. [12] and Mohan et al. [11] were the only authors to implement a procedure to calibrate the system to any arbitrary orientation of the smartphone. All of the other systems assume that the smartphone is placed in a fixed position within a vehicle. Automatic virtual reorientation of a smartphone's axes to a vehicle's axes is considered a best practice for any smartphone-based vehicle monitoring system.

III. SYSTEM DESIGN

The proposed algorithmic approach used to detect aggressive driving is discussed in this section. The hardware setup used to collect driving data with which the system was developed and tested is also described.

The vehicle's axes are denoted as x' , y' and z' in the directions as shown in Figure 2. The smartphone's axes are denoted as x pointing towards the right and y to the top from the phone's front, while z points out orthogonal to the screen. The system assumes the smartphone's axes are aligned with the vehicle's axes. Readings from the accelerometer's three axes (x, y, z) are denoted as a_x , a_y and a_z . Readings from a gyroscope's three axes are denoted as ω_x , ω_y and

ω_z . Accelerometer readings are expressed in terms of the acceleration from gravity, g (9.8 m/s^2), and gyroscope readings in terms of rotation rate (rad/s).

A. Aggressive driving model

Aggressive driving is considered as deliberate behaviour by a driver to perform any manoeuvre in such a manner that increases the risk of a road accident. Such deliberate driving behaviour often involves exceeding the speed limit.

In developing countries, such as in Africa, roads and vehicles are generally not as well maintained as in North America and Europe. Speeding is thus a bigger contributing factor to road accidents in Africa than typically elsewhere. Enforcing speed limits in rural areas proves to be difficult, because of typical budgetary constraints of law enforcement agencies in Africa. Making drivers aware of the danger of speeding has always been a top priority of road safety initiatives, such as the Arrive Alive campaign in South Africa. Drivers are unfortunately not always aware they are driving too fast for the shape of the road they are on. The goal is therefore to make a driver aware of unsafe speeds for the specific road they are on using their own smartphone.

Our aggressive driving model is consequently based on the angle of a turn, the lateral force exerted on the vehicle and its speed through the turn. The gyroscope, accelerometer and GPS of a smartphone is used accordingly to obtain the required information.

B. Recognition algorithm

For the recognition of lateral driving manoeuvres, the a_x acceleration and rotation rate ω_z are used. The accelerometer and gyroscope are continuously sampled at a rate of 20 Hz, in line with [9].

Figure 3 shows a block diagram of the system. The accelerometer output is band-pass filtered to remove sensor noise and the gravitational force vector, as its direction changes slowly when the vehicle's roll and pitch changes while driving. The gyroscope output is low-pass filtered to remove noise.

In order to detect manoeuvres, the start and end of driving events are determined by using the endpoint detection algorithm. For lateral manoeuvres, a simple moving average (SMA) of ω_z is continuously calculated over 40 samples. The beginning of a lateral event is detected if the SMA goes above a set threshold. The previous 40 and succeeding samples of ω_z are concatenated until the SMA falls below the threshold, signifying the end of the event. The samples of a_x are also saved during the same time window. An event is dismissed if it is less than 2.5 or more than 15 seconds long. This is to keep the system from hanging on potentially erroneous or noisy data. The length boundaries were established empirically to detect most valid events.

When a valid driving event has been detected, the signals recorded during the event are compared to a set of templates using the dynamic time warping (DTW) algorithm [15]. DTW finds an optimal alignment between two signal vectors with different lengths. Consider a matrix of the Euclidean distance

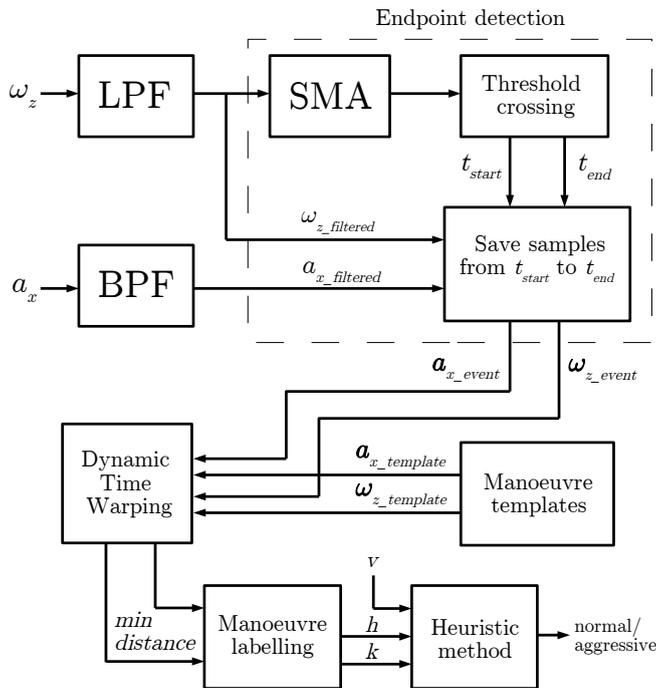


Fig. 3. Block diagram of the system.

between each point of two signal vectors, as seen in Table II. Both vectors start at the bottom left corner. An optimal warping path constitutes the minimum sum of distances, while adhering to monotonicity, boundary and step size conditions. The template with the lowest minimum-distance warp path to the detected event is the closest match.

C. Empirical classification

The acceleration and rotation rate templates are discrete Gaussian signals with fixed lengths that were created from collected driving data. The ω_z templates indicate the angle of a turn. It allows the system to classify a left or right bend from 1 to 3, based on the closest matching ω_z template — with 1 indicating an easy bend, 2 a medium bend and 3 a sharp bend. Similarly there are six a_x templates with increasing amplitudes.

TABLE II
DTW COST MATRIX SHOWING OPTIMAL WARPING PATH.

Template	Minimum-distance									0
0	0	-1	-2	-3	-4	-4	-2	-1	-1	0
1	1	0	-1	-2	-3	-3	-1	0	0	1
2	2	1	0	-1	-2	-2	0	1	1	2
4	4	3	2	1	0	0	2	3	3	4
3	3	2	1	0	-1	-1	1	2	2	3
3	3	2	1	0	-1	-1	1	2	2	3
1	1	0	-1	-2	-3	-3	-1	0	0	1
0	0	-1	-2	-3	-4	-4	-2	-1	-1	0
Measured	0	1	2	3	4	4	2	1	1	0

A heuristic method is used to label any recognized turn as taken normally or aggressively, based on the vehicle's speed (obtained from the GPS) and matching a_x and ω_z template. From experimental results it was evident that two conditions need to be satisfied to classify a turn as aggressive:

1. $v > 50(3 - h)$
2. $k > 4 \vee k > (h + 2)$

where v is the vehicle's speed in km/h, h is the labelled bend severity (1–3) and k is the a_x template number (1–6).

D. Hardware setup

A Samsung Galaxy S3 smartphone was used for driving data collection. A simple data logger Android application was developed that samples the accelerometer and gyroscope at 20 Hz. Although a higher sampling rate is possible, it increases power consumption, and 20 Hz was considered fast enough for the proposed system. The application saves the sensor samples and GPS data to an SQLite database.

In order to validate the smartphone's data, an Arduino board was used to also log data from a dedicated GPS and inertial measurement unit (IMU) to an SD card.

IV. EXPERIMENTAL RESULTS

The collected data set and tested system performance is presented in this section.

A. Data collection

Six individuals were asked to drive a pre-determined route while subjective labelling of their turns were performed by hand. A route of 15 km was chosen that has varying bends and up- and downhill parts. The route necessitates drivers to manage their speed as straighter sections are followed by several sharp bends. All the distinct bends were annotated by hand on a map with a severity of 1, 2 or 3. The route has 55 identified bends — 28 right and 27 left bends. Route notes were used to label how the driver took each bend: normally or aggressively. Although the labelling was done subjectively, it was kept consistent for each driver.

The raw data was post-processed and valid data was successfully extracted and labelled for 387 bends. Overall, the endpoint detection algorithm successfully detected 95% of the left and right bends. The data was split in a 66%/33% ratio for training and test data respectively. The training data set was used to create gyroscope and accelerometer signal templates for the three bend severities taken both normally and aggressively. Twelve templates were thus created from the gyroscope and accelerometer data in total.

B. System performance

The test data set was used to obtain the results given in Table III. For the driver labelled as most aggressive from first-hand observation, the classifier achieved a FN and FP rate of 80% and 10.5%, respectively. Figure 4(a) shows the lateral acceleration of a one minute section where 4 of his aggressive turns occurred. The vehicle's average speed was 85 km/h through this section.

TABLE III
 CLASSIFICATION RESULTS.

Bend severity classification:						
Accuracy	= 83.7%					
Aggressive manoeuvre classification:						
Precision	= 64.3%					
Recall	= 37.5%					
Specifity	= 95.2%					
Accuracy	= 84.5%					
		TP	FP	=	9	5
		FN	TN		15	100

Figure 4(b) shows the lateral acceleration of another driver for the same section of road as in Figure 4(a). All of the second driver's turns were observationally labelled as normal, and his average speed was 70 km/h through the section. The lateral acceleration never exceeded $0.1g$, whereas with the aggressive driver the acceleration exceeded $0.1g$ for all four turns. The classifier achieved a FN and FP rate of 0% and 5.9%, respectively, for this driver.

With 24 aggressive turns out of 129 in the test set, the aggressive turn labelling heuristic achieved a FN and FP rate of 62.5% and 4.8%, respectively. Although the FN rate is high, a lower FP rate is desirable. It is biased to label a driver as aggressive based on falsely identified aggressive manoeuvres. The heuristic was tuned to obtain the least false positives, at the expense of missing many true positives (TP). Although the sample size was small, it is clear that the classifier's precision and recall is poor and could be improved. The strength of the system is that it can definitely be expanded to recognize other manoeuvres by preparing relevant templates for the same or other axes of the sensors.

V. CONCLUSIONS

This paper presents a driving manoeuvre recognition and classification system that is suitable for implementation on

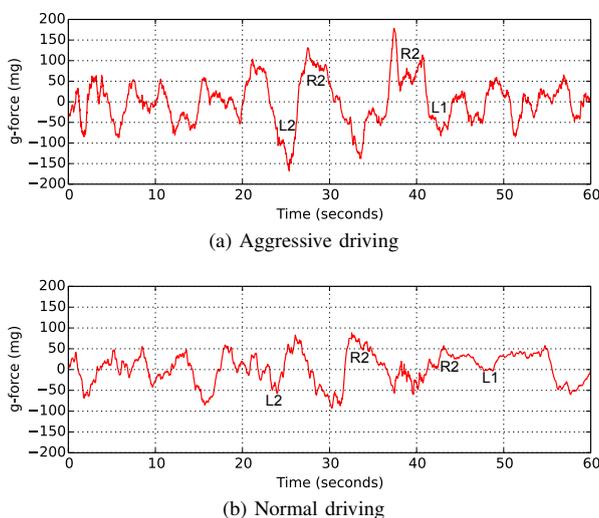


Fig. 4. Band-pass filtered lateral acceleration of left (L) and right (R) turns labelled with the bend severity (1–3).

a smartphone. The recognition algorithm can successfully detect turns of varying severity by comparing the gyroscope signal to a template set, using dynamic time warping. The system can also label each recognized turn as taken normally or aggressively by the driver. The system can be expanded to recognize a variety of manoeuvres. Such a system could be used to monitor a driver over a long period and give him feedback on how to drive safely. The prevalence of smartphones also allows such a system to be easily and cost-effectively deployed on a large scale. In future work we will compare the accuracy of the proposed manoeuvre classification approach to that of other machine learning techniques.

ACKNOWLEDGEMENT

The authors would like to thank MTN for their continued financial support through the MTN Mobile Intelligence Lab.

REFERENCES

- [1] "Global status report on road safety: time for action," Geneva, World Health Organization, 2009, [ONLINE] Available: http://whqlibdoc.who.int/publications/2009/9789241563840_eng.pdf. [Accessed 28 August 2014].
- [2] "Aggressive driving: Research update," AAA Foundation for Traffic Safety, April 2009, [ONLINE] Available: <http://www.aaafoundation.org/pdf/AggressiveDrivingResearchUpdate2009.pdf>. [Accessed 28 August 2014].
- [3] "DriveCam programs," Lytx, 2014, [ONLINE] Available at: <http://www.lytx.com/our-solutions/drivecam-programs>. [Accessed 28 August 2014].
- [4] "How it works," AutoHabits, 2012, [ONLINE] Available at: <http://autohabits.com/how-it-works>. [Accessed 28 August 2014].
- [5] "Fleet safety," FleetMind, 2014, [ONLINE] Available at: <http://www.fleetmind.com/fleet-management-products/fleet-safety>. [Accessed 28 August 2014].
- [6] R. Xu, S. Zhou, and W. J. Li, "MEMS accelerometer based nonspecific-user hand gesture recognition," *IEEE Sensors Journal*, vol. 12, no. 5, pp. 1166–1173, 2012.
- [7] A. Akl, C. Feng, and S. Valaee, "A novel accelerometer-based gesture recognition system," *IEEE Transactions on Signal Processing*, vol. 59, no. 12, pp. 6197–6205, 2011.
- [8] S. Reddy, M. Mun, J. Burke, D. Estrin, M. Hansen, and M. Srivastava, "Using mobile phones to determine transportation modes," *ACM Transactions on Sensor Networks (TOSN)*, vol. 6, no. 2, p. 13, 2010.
- [9] D. A. Johnson and M. M. Trivedi, "Driving style recognition using a smartphone as a sensor platform," in *14th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2011, pp. 1609–1615.
- [10] J. White, C. Thompson, H. Turner, B. Dougherty, and D. C. Schmidt, "WreckWatch: automatic traffic accident detection and notification with smartphones," *Mobile Networks and Applications*, vol. 16, no. 3, pp. 285–303, 2011.
- [11] P. Mohan, V. N. Padmanabhan, and R. Ramjee, "Nericell: rich monitoring of road and traffic conditions using mobile smartphones," in *Proceedings of the 6th ACM conference on Embedded Network Sensor Systems*. ACM, 2008, pp. 323–336.
- [12] J. Dai, J. Teng, X. Bai, Z. Shen, and D. Xuan, "Mobile phone based drunk driving detection," in *4th International Conference on Pervasive Computing Technologies for Healthcare (PervasiveHealth)*. IEEE, 2010, pp. 1–8.
- [13] H. Eren, S. Makinist, E. Akin, and A. Yilmaz, "Estimating driving behavior by a smartphone," in *Intelligent Vehicles Symposium (IV)*. IEEE, 2012, pp. 234–239.
- [14] M. Fazeen, B. Gozick, R. Dantu, M. Bhukhiya, and M. C. González, "Safe driving using mobile phones," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 3, pp. 1462–1468, 2012.
- [15] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561–580, 2007.

Automated Landing of an Intelligent Unmanned Aerial Vehicle in Crosswind Conditions using Total Energy Control

C.T. Le Roux

Department of Electrical & Electronic Engineering
Stellenbosch University
Stellenbosch, South Africa
Email: cornelusleroux@gmail.com

J.A.A. Engelbrecht

Department of Electrical & Electronic Engineering
Stellenbosch University
Stellenbosch, South Africa
Email: jengelbr@sun.ac.za

Abstract—This paper presents the development, implementation and verification of a flight control system for the automated landing of an intelligent unmanned aerial vehicle (UAV) in crosswind conditions.

There is an increasing number of commercial opportunities for UAVs in business, agriculture, industry and mining, the emergency services and security services. The major barrier to commercialisation of UAVs is the certification process, where automated take-off and landing is a key feature required.

The automated landing system presented in this paper uses a longitudinal control system based on the total energy control system (TECS), and a lateral control system that combines a heading and guidance controller with a cross-track error controller. A software state machine is used to advance the flight control system through the different stages of the automated landing. The TECS architecture allows the airspeed and flight path angle to be decoupled, while the Cross-Track Controller uses a limited integrator to drive the cross-track error to zero in the presence of crosswind.

The automated landing system is implemented on a UAV with an on-board computer, sensors and actuators, and is verified in simulation and with practical flight tests. The hardware simulation results show that the UAV is able to land autonomously in crosswinds up to 3.6 metres per second, with a landing accuracy of 3.50 metre in-track and 0.12 metre cross-track.

I. INTRODUCTION

There is an increasing number of commercial opportunities for UAVs in business (aerial photography, speed courier services), agriculture (surveying, crop inspection, crop dusting, farm security), industry and mining (power line inspection, prospecting), the emergency services (disaster monitoring, delivery of emergency supplies, firefighting) and security services (surveillance, policing). However, a major barrier to the commercialisation of unmanned aircraft is the certification process. Before UAVs can be operated in civil airspace, they must first pass a rigorous certification process to prove that they will operate safely. One of the key enabling technologies required for the certification and eventual integration of intelligent unmanned aircraft into commercial airspace is automated take-off and landing.

Autonomous landing systems are currently researched globally by various institutions for different applications. Cho

et al. [4] developed a system using only a single-antenna GPS receiver, implementing differential GPS (DGPS) for increased accuracy in position information. The only extra sensor used was for airspeed via the pitot tube, as accurate airspeed measurements are very important during the landing phase. López et al. [8] presented a paper on the differences between H_∞ and quantitative feedback theory techniques that should be robust against wind disturbances and control the altitude accurately. They found that controllers designed by both techniques guaranteed robust stability and attenuated high frequency noise due to sensors supplying suitable control signals. Masuko et al. [9] opted to use visual feedback, using a small Linux ARM computer running OpenCV. The high velocity of the aircraft resulted in blurred images taken by the camera which could not be processed fast enough to ensure safe landing.

Akmeiliawati and Mareels [1] presented a non-linear energy-base control method (NEM) based on passivity-based control techniques similar to TECS. The difference between their technique from TECS was that the non-linearity of the system dynamics were directly taken into account, where the aircraft dynamics are expressed in Euler-Lagrange equations of motion derived from the energy equations. In [2], they extended the NEM technique by further exploiting the inherent time scales of the dynamics using a singular perturbation technique to simplify the overall design. The aircraft is treated as a single point mass while disregarding the fast pitch and elevator dynamics. The system conforms to the Lyapunov stability criteria, and good stability and performance were achieved during Monte Carlo simulations. Looye and Joos [7] used multi-objective optimisation to design a controller with the purpose to synthesise the free parameters (gains, time-constants) in these controller functions by using parameter weighting, sequentially expanded to the simultaneous optimisation of all functions. The system was successfully flight tested, however the glide slope and disturbance rejection criteria did not work to full satisfaction.

This paper focus on a simpler design using accurate sensors, mainly the NovAtel DGPS system operating in ALIGNTM mode to provide very accurate position measurements. The



Fig. 1. Photograph of a Phoenix Trainer 60, the model aircraft used in this project.

simpler design introduces fewer modelling errors and speeds up the design process and testing iterations while keeping functionality. The end goal of the project is to make the UAV capable of landing on a moving platform, such as a naval vessel.

II. MATHEMATICAL MODELS

To design and test the system in simulation, a model is required representative of the aircraft, avionics, and environmental factors. The models used are discussed below.

A. Aircraft

The aircraft model used is explained thoroughly in [5] using the standard North-East-Down convention for inertial space and Euler-3-2-1 angle conversions. The model includes

- 1) forces and moments, which include the aerodynamic, engine/thrust, and gravitational models; and
- 2) six degrees of freedom equations of motion, which includes the kinetic and kinematic models.

The model aircraft used is a Phoenix Trainer 60, modified with custom sensors and actuators. The aircraft will be flown by the autopilot system, but a safety pilot can assume control if required. An illustration of the vehicle is shown in Fig. 1.

B. Wind Model

With the focus on landing during windy conditions, the Dryden and Von Kármán wind models, as presented in military standards [10] and [11], were implemented in simulation to more accurately reflect wind conditions and the effect thereof on the aircraft. These include models for

- 1) turbulence: the irregular forces and moments acting on the aircraft caused by chaotic winds;
- 2) shear: the variation in wind velocity caused by difference in altitude; and
- 3) gusts: shorts bursts of high velocity wind.

C. Sensor Models

To reflect inaccuracies in the on-board measurements systems, models of the sensors used on-board the model aircraft are implemented to include noise on the measurements they provide. These include models for

- 1) 3-axis magnetometer;

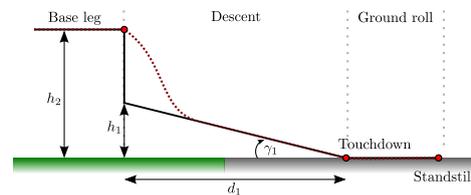


Fig. 2. Side view of the flight path during the different states of landing.

- 2) 3-axis gyroscope;
- 3) 3-axis accelerometer;
- 4) global positioning system (GPS); and
- 5) pressure sensor (pitot tube).

III. CONTROL SYSTEM DESIGN

The design of the controllers has seen continuous adaptation throughout the project. The controllers implemented are explained in their respective sections below.

A. Landing Controller

The airfield where the system will be flight tested has obstacles in the runway path that may threaten safe flying conditions. To circumvent this, a flight path is proposed as illustrated in Fig. 2. After flying at $h_2 = 20$ m altitude until sufficient obstacle clearance is achieved, a ramp starting at a lower altitude will be commanded at a flight path angle of $\gamma_1 = 3.5^\circ$ until the touchdown point. This sequence is activated $d_1 = 200$ m before the touchdown point so that the ramp initial altitude is $h_1 = 12.23$ m. This method is proposed because the aircraft cannot be trimmed to a much steeper angle while keeping low airspeed. The aircraft is thus allowed to reduce altitude as fast as possible to reach the region in which it can be successfully trimmed to the landing conditions. Since the end goal of the project is landing on a moving platform, the flare procedure is omitted and the landing executed similarly to real pilot touchdowns when landing on naval vessels. Upon touchdown, a safety pilot will assume control of the aircraft and bring it to a halt as runway taxiing is outside the scope of this project.

The trim airspeed was chosen at 16 m/s, which is well above the aircraft's stall speed of approximately 12 m/s and therefore deemed safe. The low glide slope may reduce landing accuracy as shown in predecessors of this project, [3] and [13].

Two other methods are also proposed to apply course correction for both a stationary and moving touchdown point. The first uses proportional navigation techniques to supply longitudinal reference signals to regulate the line-of-sight angle between the aircraft and the touchdown point. The other uses touchdown point estimation and trigger re-planning of the glide path when the error would be out of predefined bounds. As the system is still under development, these techniques will not be discussed further in this paper.

B. Longitudinal Controller

The concept of an energy controller was pioneered and patented by [6] in 1985. The main purpose of this controller is

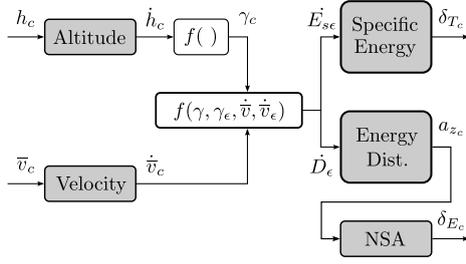


Fig. 3. Simplified Total Energy Control System concept with the pitch control inner loops replaced by a Normal Specific Acceleration Controller.

to balance the kinetic and potential energy of the aircraft with an integrated and simplified control system architecture, using the throttle to control the aircraft's total energy error and the elevator to control the energy distribution error between flight path and airspeed. The explanation below is illustrated by the simplified block diagram of the TECS architecture in Fig. 3, with all parameters designed in a heuristic fashion.

The aircraft's total energy E can be expressed by

$$E = mgh + \frac{1}{2}mV^2 \quad (1)$$

which is the sum of the potential and kinetic energy components, where m is the aircraft's mass, g is the gravitational acceleration, h is the altitude of the aircraft, and V is the velocity. By assuming a constant weight of $W = mg$, Eq. 1 can be rewritten and time-derived to form

$$\dot{E} = W \left(\dot{h} + \frac{V\dot{V}}{g} \right) \quad (2)$$

Using the small angle approximation, the flight path angle γ can be calculated as

$$\gamma = \frac{\dot{h}}{V} \quad (3)$$

Substituting Eq. 3 into Eq. 2 and scaling the result by V , a velocity normalised energy rate equation is obtained, yielding

$$\frac{\dot{E}}{V} = W \left(\gamma + \frac{\dot{V}}{g} \right) \quad (4)$$

This reveals that, at a given airspeed, the rate of change of the aircraft's energy is dependant only on the flight path angle and longitudinal acceleration.

The second law of Newton, $F = ma$, is used to express the longitudinal motion of the aircraft as

$$\frac{W}{g}\dot{V} = T - D + W \sin \gamma \quad (5)$$

where T is the total thrust applied and D is the total drag. By again assuming a small flight path angle, the equation can be rewritten as

$$W \left(\gamma + \frac{\dot{V}}{g} \right) = T - D \quad (6)$$

This resembles Eq. 4, concluding that the rate of change of the aircraft's energy is proportional to the difference between thrust and drag. The required thrust can then be written as

$$T_{req} = W \left(\gamma + \frac{\dot{V}}{g} \right) + D \quad (7)$$

$$= \frac{\dot{E}_s}{V} + D \quad (8)$$

where \dot{E}_s is the specific energy rate of the aircraft.

At a specific thrust level, it is possible to exchange flight path angle and acceleration by only using the elevator. To drive the aircraft's current flight path angle and longitudinal acceleration towards desired reference values, it becomes apparent that a flight path and speed control concept is obtained in the form of total specific energy rate, given as

$$\dot{E}_{s\epsilon} = \gamma_\epsilon + \frac{\dot{V}_\epsilon}{g} \quad (9)$$

where subscript ϵ denotes the error between the reference and measured signal values. The elevator is to be driven until the energy rate distribution error

$$\dot{D}_\epsilon = -\gamma_\epsilon + \frac{\dot{V}_\epsilon}{g} \quad (10)$$

relative to the target flight path and acceleration is zero. The specific energy rate error is used directly in the computation of the thrust command δT , and the energy rate distribution error is similarly applied to the elevator command δE .

In most cases, it is desired to command the aircraft altitude and airspeed rather than the flight path angle and acceleration. To realise this, two outer loops of both proportional control are implemented in the design with gains K_h and K_v for altitude and airspeed which produce climb rate and acceleration commands, respectively. The climb rate command can also be given directly instead of being determined from the altitude controller, resulting in three operational modes to change aircraft altitude—flight path angle, climb rate or direct altitude commands. The gains K_v and K_h are selected to have equal values to provide identical altitude and airspeed dynamics. In essence, they determine the error decay time constants in the altitude and airspeed responses. Using these control mechanisms, the altitude and speed errors are now scaled in relative energy terms.

In the classic TECS, the result of the energy distribution rate controller is used as a pitch angle command, which feeds a Pitch Angle Controller with proportional gain K_θ . This in turn feeds a Pitch Rate Damper with proportional gain $K_{\dot{\theta}}$, which finally produces the elevator command. In addition to the default TECS, a climb rate feed forward was added in the climb rate loop to ensure that the correct sink rate was maintained upon landing. The Pitch Angle Controller and Pitch Rate Damper loops were replaced by a Normal Specific Acceleration (NSA) Controller which more directly controls the angle of attack of the aircraft. The proportional segments of the controllers were also fed from the error signals rather

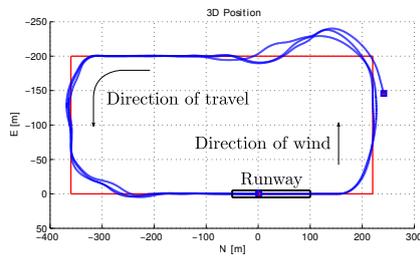


Fig. 5. Dual-mode lateral controller response on the intended waypoint track.

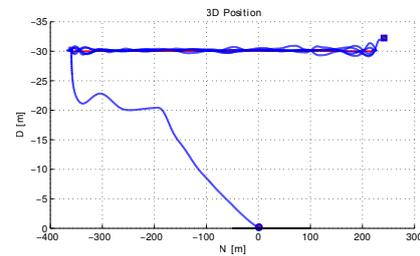


Fig. 6. TECS altitude regulation and landing sequence.

In the simulation, the on-board computer and extended Kalman filter are initialised on the touchdown point on the runway, after which a manual take-off is performed. The autopilot is activated and go-arounds are performed to test the regulation of altitude, airspeed and track navigation. After sufficient go-arounds, the landing command is given and the aircraft continues to fly the circuit with the landing sequence started when the final waypoint and approach points are reached. The aircraft follows the landing states and touches down on the runway as illustrated in Figs. 5–6. It was subjected to turbulence and an east-to-west crosswind of 10 km/h, which is over 17% of the flight speed.

TECS regulates the altitude to within a 1.0 m error, barely losing altitude when rolling at angles of up to 30°. However it does increase the airspeed by up to 10%, exchanging potential energy for kinetic energy. For the straight legs, it can be seen that the cross-track error is reduced to zero shortly after the system is switched from the Heading and Guidance to the Cross-Track Error Controller. The early waypoint switching method works to satisfaction, providing a better transient response similar to the non-linear guidance method as proposed by Park [12] and implemented by Alberts [3]. A noticeable effect is observed when the aircraft is subjected to a tailwind, which causes loss of lift force and increases waypoint track overshoot.

The simulations showed 95% confidence in landing within 3.50 m in-track and 0.12 m cross-track, which is 1.75 and 0.06 wingspan, respectively. Although the longitudinal accuracy is good for landing on a fixed runway or large naval vessel, it still needs improvement to obtain an accuracy better than 0.5 wingspan to land on a moving platform during a practical flight test. To improve the longitudinal accuracy, the landing strategy will be revised and complemented by proportional navigation techniques. The lateral accuracy is exceptional for such a small aircraft in light wind conditions and do not need any further improvement.

V. CONCLUSION

This paper presented the implementation of concepts that would allow a UAV to successfully land with acceptable accuracy in crosswind conditions using total energy control and an aggressive cross-track control system.

A modified landing sequence, controlled by a software state machine, was presented to allow for a flight-testable landing

on a runway with hazards. Longitudinal performance in TECS shows good control during the circuit navigation and landing phases, even with high roll angle references given by the lateral controllers. Upon landing, the system does achieve the correct sink rate as well as a positive pitch angle. Lateral performance with the combination of the Heading and Guidance Controllers and the Cross-Track Controller can be considered exceptional and further improvements are not required, as the landing error is almost zero under acceptable disturbances.

Further work would include attempting to increase longitudinal accuracy by incorporating proportional navigation or estimation and re-planning methods. This is not only beneficial to a runway landing, but also to a landing where the touchdown point is moving. The system is currently being subjected to further HIL simulations in preparation for flight tests, which should yield more results on practical feasibility.

REFERENCES

- [1] R. Akmeliawati, I. Mareels, *Nonlinear Energy-based Control Method for Aircraft Dynamics*. Decision and Control, 2001. Proceedings of the 40th IEEE Conference on, Pages 658–663, 2011.
- [2] R. Akmeliawati, I. Mareels, *Nonlinear Energy-Based Control Method for Aircraft Automatic Landing Systems*. Control Systems Technology, IEEE Transactions on, Vol. 18, Pages 871–884, 2010.
- [3] F. N. Alberts, *Accurate Autonomous Landing of a Fixed-Wing Unmanned Aerial Vehicle*. Stellenbosch University, 2012.
- [4] A. Cho, J. Kim, S. Lee, S. Choi, B. Lee, B. Kim, N. Park, D. Kim, C. Kee, *Fully Automatic Taxiing, Takeoff and Landing of a UAV Using a Single-Antenna GPS Receiver Only*. Control, Automation and Systems, 2007. International Conference on, Pages 821–825, 2007.
- [5] M. V. Cook, *Flight Dynamics Principles*, 3rd ed. Elsevier Ltd., 2007.
- [6] A. A. Lambregts, *Total Energy Based Flight Control System*, US Patent Nr. 6062513. The Boeing Company (Seattle, WA), 1985.
- [7] G. Looye, H.D. Joos, *Design of Autoland Controller Functions with Multi-Objective Optimization*. AIAA Guidance, Navigation, and Control Conference and Exhibit, 2002.
- [8] J. Lopéz, R. Dormiro, J. P. Gómez, *A Fully Autonomous UAV Landing Controller Synthesis - QFT and H_∞ Technique Comparison*, 3rd ed. Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering 2012, Pages 281–293, 2011.
- [9] K. Masuko, I. Takahashi, S. Ogawa, W. Meng-Hung, A. Oosedo, T. Matsumoto, K. Go, F. Sugai, A. Konno, M. Uchiyama, *Autonomous Takeoff and Landing of an Unmanned Aerial Vehicle*. System Integration, International Symposium on, Pages 248–253, 2010.
- [10] United States Department of Defense, *MIL-HDBK-1797*. United States Department of Defense, 1997.
- [11] United States Department of Defense, *MIL-F-8785C*. United States Department of Defense, 1980.
- [12] S. Park, J. Deyst, J. P. How, *A New Nonlinear Guidance Logic for Trajectory Tracking*. In AIAA Guidance, Navigation, and Control Conference and Exhibit, Pages 1–16, 2004.
- [13] S. J. A. Smit, *Autonomous Landing of a Fixed-Wing Unmanned Aerial Vehicle using Differential GPS*. Stellenbosch University, 2013.

Performance Measurement Trends in the implementation of Intelligent Transportation Systems (ITS) within the South African Transportation Environment

C. B. Struwig and S. J. Andersen, Stellenbosch Smart Mobility Lab (SSML) and Department of Civil Engineering, *Stellenbosch University, South Africa**

Abstract – Over the past decade, the South African transportation environment has actively started to adapt a technology-driven setting. Intelligent Transportation Systems (ITS) applications such as Advanced Traffic Management Systems (ATMS) and Advanced Public Transportation Systems (APTS) have since been promoted and developed. These ITS deployments have brought about new areas for consideration. If the sustainability of the newly developed systems is to be ensured, sufficient attention needs to be given to the managing of their inherent technology-related aspects. These aspects are currently, to varying degrees, being measured and monitored. However, little thought is given to ITS performance management in the conceptualization- and planning phase of ITS projects. As a result, the monitoring is mostly done by a modular- and possibly inconsistent performance measurement approach. Moreover, in the absence of a set of widely accepted performance measures and transferable methodologies, it is very difficult for the local industry to objectively assess the effects of their specific applications with regard to the implementation of policies and technologies. The aim of this paper is thus to define a common evaluation framework for the monitoring and managing of the newly developed systems and to present guidelines as to its application. The aforementioned is accomplished by elucidating the need for managing performance measurement and by providing a review on the current ITS measurement trends and movements in the South African transportation environment. Ultimately, a major evolution in the nation's transportation environment - in the form of an ITS performance management regime - may be stimulated.

I. FOREWORD

This paper is of an investigative nature and is the first in a series of publications emanating from a research project on ITS performance management. The specific application environment addressed is that of a developing country such as South Africa (SA). In this paper, the need for managing performance measurement is elucidated and a review on the current ITS measurement trends is provided. The recommended way forward and the future directions are also discussed.

* This research is based on a thesis presented in fulfillment of the requirements for the degree of Doctor of Philosophy in Civil Engineering and is supported by Stellenbosch University, South Africa.

Miss C. B. Struwig is a full-time PhD student at the Department of Civil Engineering at Stellenbosch University busy conducting her research in ITS (email: cstruwig@sun.ac.za).

Dr. S. J. Andersen is an Industry Associate Professor in ITS at the Department of Civil Engineering at Stellenbosch University (email: jandersen@sun.ac.za).

II. BACKGROUND

In 2007 the South African Government introduced the National Land Transportation Act (NLTA) in which the minimum requirements for the nation's Integrated Transportation Plans (ITPs) are stipulated [6]. The provision of the NLTA has given rise to a significant increase in implementation mandates; especially with regard to the deployment of ITS applications such as ATMS and APTS. In general, SA has started to adapt a technology-driven setting where mobility for all, system interoperability and seamless traveling are fostered.

III. INTRODUCTION

A. Problem Formulation

Since ATMS and APTS are examples of ITS applications, the critical role that technology plays in their realization and operation is evident. For example, the implementation of Freeway Management Systems (FMS) is dependent on technology such as traffic control devices and the implementation of Integrated Rapid Transportation (IRT) systems is dependent on technology such as electronic payment devices. These inherent technology-related aspects of the ITS applications need to be managed.

These aspects are currently, to varying degrees, being measured and monitored. However, little thought is given to ITS performance management in the conceptualization- and planning phase of ITS projects. As a result, the monitoring is mostly done by a modular- and possibly inconsistent performance measurement approach. Moreover, the resulting required multi-facet deployment associated with implementing such technology-oriented systems renders the ease of their management. Consequently, SA currently has no (holistic) performance management approach for measuring the technology-related aspects of ITS projects. Given this general lack of an established approach for a performance management regime as well as the absence of management tools which can allow for the regular- and consistent measurement of such systems' performance, no real conclusions can be drawn to make an informed decision about the existing systems' overall health.

This current state of affairs creates skepticism around the sustainability of the newly deployed ITS applications. Without consistent, pre-determined and pre-specified standards to measure their performance, the degradation of the systems (over time) is highly probable. This likelihood of degradation is also further exacerbated by the ever changing world of technology we find ourselves in.

B. Proposition and Motivation

In order to ensure the sustainable deployment of ITS applications, their technology-related aspects need to be managed in a pre-defined, continuous and holistic manner. In essence, the local industry needs to be able to regularly and objectively assess the effects of their specific applications with regard to the implementation of policies and technologies. This gives rise to the adoption of a performance management regime. However, for such a regime to be employed, a performance measurement framework first needs to be developed. This framework will serve as the reference point for ITS performance management by presenting guidelines to widely accepted performance measures and by capturing transferable methodologies. If such a framework is in place, implementing agencies will be able to obtain the necessary knowledge to easily make day-to-day informed decisions regarding the overall performance of their respective systems and decision makers will be provided with a useful evaluation tool to aid in the continuous assessment of their investment in transportation technology.

C. Focus of Paper

The emphasis of a performance measurement framework is to act as a management tool for assessing the effective development and maintenance of ITS applications. The aim of this paper is thus to define an evaluation framework for the monitoring and managing of the technology-related aspects inherent to ATMS and APTS. The aforementioned is accomplished by elucidating the need for managing performance measurement and by providing a review on the current ITS measurement trends and movements with regard to technology deployments. Through the analysis of the status quo, various shortages are identified. These shortages assist in filling the identified gaps and aid in identifying measurement guidelines which then contribute towards the establishment of the performance measurement framework.

IV. MANAGING PERFORMANCE MEASUREMENT

In order to comprehend the concept of managing performance measurement, firstly an understanding of performance **measurement** and then, secondly, an understanding of performance **management** are needed.

A. Performance Measurement

According to [3], performance measurement is defined as the assessment of an organization's output as a product of the management of its internal resources (e.g. money, people, vehicles and facilities) and the environment in which it operates. Therefore, in general terms, performance measurement refers to any evaluation- or comparison measure. These measures can either be a quantitative- or a qualitative characterization of performance; with each measure having certain indicators that are used to signify the performance of the system under consideration.

As can be seen in Fig. 1, the performance measurement process can be disassembled into a systematic top-down approach.

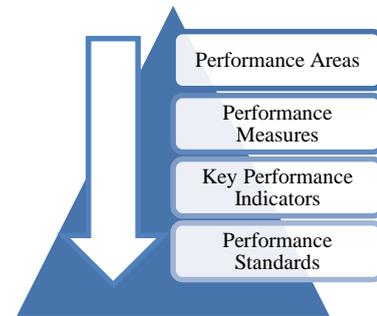


Figure 1. A Systematic (top-down) Approach towards Performance Measurement

B. Performance Management

As stated in [1], performance-based management is a systematic approach to performance improvement through an ongoing process of establishing strategic performance objectives; measuring performance; collecting, analyzing, reviewing and reporting performance data; and using that data to drive performance improvement. In essence, performance measurement is thus a critical component and predecessor of performance-based management.

The performance management process can be subdivided into six main sub-processes. These are [1]:

1. Establish a performance-based management program.
2. Establish an integrated performance measurement system.
3. Establish accountability for performance.
4. Collect data to assess performance.
5. Analyze, review and report performance data.
6. Use performance information to drive improvement.

V. STATUS QUO

As mentioned previously, no consistent or holistic performance management approach for measuring the technology-related aspects of the ITS deployments in the South African transportation environment is currently available. In order to investigate and elucidate the need for the provision of a performance management approach, a typical transportation environment that reflects the status of the current technology implementation with regard to the nation's ITS applications should be considered. An example of such a representative transportation environment is the City of Cape Town (CoCT).

As a point of reference, the CoCT's current transit- and traffic operations, including background to the extensive recent implementation of technology systems in this transportation environment, are considered. A discussion of these systems and their relating ITS-aspects follows in the next sub-sections.

A. Technological Developments and Applications

With the evolvement of information technology, immense scope for growth in the utilization of information systems has been created. Several initiatives within the transportation

industry serve as testimony of this fact. A discussion of some of the evident emerging technology developments and their applications follows.

1) *The Transportation Management Center*

In May 2010, the CoCT officially opened its Transportation Management Center (TMC). The TMC is the City's operations' facility, with resultant transportation data repository, that is regarded as one of the finest state-of-the-art multi-functional facilities in the world [2]. Since the realization of this center, extensive deployment of technology and the relating supporting ITS devices have been implemented. The TMC has five main functional areas, namely: 1) FMS, 2) Arterial Management System: AMS/Urban Traffic Control: UTC, 3) Integrated Incident Management: IIM, 4) IRT and 5) Transportation Information Center: TIC [3].

In Fig. 2, the general transit- and traffic operations of the TMC are portrayed. These operations are presented under the groups of: 1) input, 2) information processing, 3) action and 4) output.

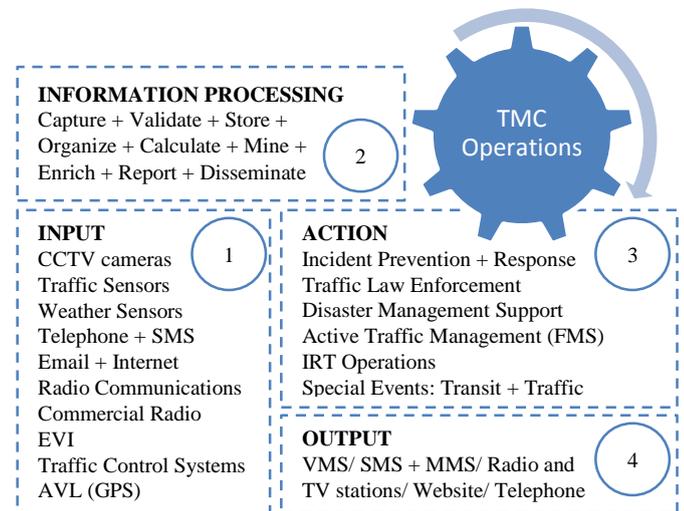


Figure 2. TMC Operations

The information retrieved from the real time monitoring is portrayed to the commuters on output devices such as Passenger Information Display Systems (PIDS) at stations and Advanced Traffic Information Systems (ATIS).

These APTS procedures facilitate IRT and are also currently managed by the CoCT's TMC. Commuters are presently informed about the arrival- and departure status (and overall timeliness) of buses and trains. Moreover, the CoCT's IRT network (referred to as MyCiti) implements Automated Fare Collection (AFC) by using an EFP system in conjunction with a smart card interoperable fare media type.

2) *A Shift towards a Technology-oriented Approach*

a) *Advanced Traffic Management Systems*

ATMS utilize ITS functions such as: traffic control systems, lane- and incident management, ramp metering, navigation- and warning systems, adaptive signal management and electronic toll collection. These ITS functions rely on information technologies in order to connect traffic sensors and roadside equipment, vehicle probes, Closed Circuit Television (CCTV) cameras, Electronic Vehicle Identification (EVI) and other devices together to create an integrated view of traffic flow and to detect accidents, dangerous weather events or other roadway hazards [4]. The information retrieved from the real time traffic monitoring is portrayed to drivers on output devices such as Variable Message Signs (VMS).

These ATMS procedures are currently managed by the CoCT's TMC and assist with FMS, AMS, UTC and IIM. Drivers are presently informed about the roadway performance and the extent and duration of incidents. Moreover, within the near future, when traffic detectors have been implemented, real time travel time data will also be available for certain freeway segments.

b) *Advanced Public Transportation Systems*

APTS utilize ITS functions such as: electronic ticketing, navigation- and warning systems, parking guidance, fleet management, cruise control and priority systems. These ITS functions rely on information technologies in order to connect Global Positioning Systems (GPS) and Automated Vehicle Location (AVL), Electronic Fare Payment (EFP) and other devices together to create a real time view of the status of all the assets and the movement of the commuters in the public transportation system [4].

B. *Performance Measurement and -Management*

1) *Private Transportation Environment*

As part of the procurement of a national ITS framework, the Government funded agency SANRAL (South African National Roads Agency Limited) has developed a measurement framework for managing contract performance.

a) *SANRAL's Contract Performance Measurement*

SANRAL has pursued a Key Performance Indicator (KPI) approach towards establishing the performance of the FMS in Gauteng, Kwazulu-Natal and the Western Cape [7]. Their primary objective is to deliver both regionally- and nationally integrated ITS functions at a consistent high level of quality. The developed performance measurement system encapsulates the following components: 1) employer's requirements, 2) principles of governance, 3) performance measurement and 4) payment mechanism. For the purpose required herein, only component three will be considered in more detail.

The concept of SANRAL's KPI approach towards performance measurement can, essentially, be represented through Fig. 3. This figure is adapted from the work in [7]. As can be seen in Fig. 3, SANRAL's measurement framework inherently caters for strategic alignment. They aim to align, from the beginning to the end of the

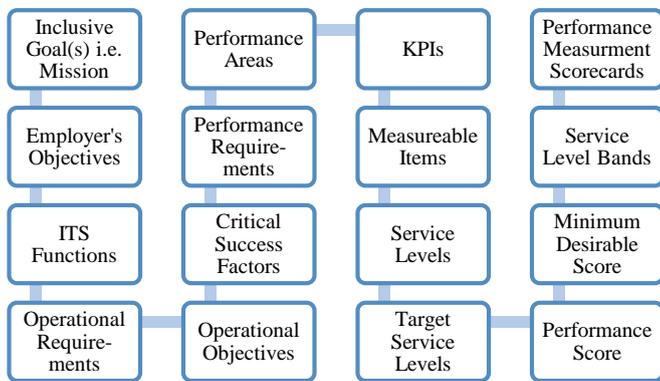


Figure 3. SANRALs Contract Performance Measurement

performance measurement contract, the operations of the main contractor with their objectives.

Unfortunately, since full functionality has not yet been deployed in the Western Cape, only certain KPIs in some of the identified performance areas are being measured. These areas include: 1) system availability, 2) incident responsiveness and information dissemination and 3) contract performance management. Moreover, the measurement is done on an ad-hoc basis and is not yet fully automated. It is believed that future functionality will enable the necessary automation. Nevertheless, it needs to be noted that the exhaustive framework behind the measurement process is viewed by some transportation role-players as being too complex and perhaps even tedious.

2) Public Transportation Environment

In the Western Cape, two approaches are currently available to assist in managing the performance of the public transportation system. While the first one discussed is privately-owned, the second one is funded by the Government.

a) *WhereIsMyTransport*

The newly developed *WhereIsMyTransport* application is an all-encompassing web-based platform that spans across and caters for three interest groups. These are: 1) the operator, 2) the commuter and 3) the advertiser.

Based on the purpose of this paper, the aspect of the *WhereIsMyTransport* application that is deemed most important is the operational environment toolset provided to the operator. Operators can currently, in real time, manage their fleets, assign routes, create schedules and monitor their drivers. The *WhereIsMyTransport* platform thus caters for vehicle tracking, asset protection, fleet management as well as personnel management. In essence, this toolset ensures safety and security while improving productivity and bringing discipline. However, being at the early stages of deployment, full functionality has not yet been attained.

b) *MyCiti's Performance Measurement Toolbox*

The performance of the MyCiti system is, to varying degrees, being measured and monitored. A review of the

current measurement tools used by the MyCiti system can be found in Table 1. This table has been composed with the help of the CoCT's IRT department. A discussion of each of the performance areas listed in Table 1 follows:

1. System planning tool used to create the (optimized) base model.
2. Controlling software used to monitor real time vehicle movement.
3. Real time analyzing software used to review and (re)optimize system with regard to schedule adherence.
4. Business Intelligence (BI) objects used from a cost cutting perspective for post-analytic purposes (e.g. conducting trend analysis and reviewing routes).
5. Use onboard validators and turnstiles at stations to provide data on the number of taps/stop/route as well as load data and origin-destination data.
6. Use equipment to monitor inter alia harsh braking, swirling, aggressive acceleration and sharp cornering with the aim of reducing risky driver behavior.
7. Real time workforce, workflow, fault reporting and Service Level Agreement (SLA) monitoring system.
8. Market the MyCiti service to maintain a favorable public image.
9. Attend to inquiries, queries, complaints and compliments.
10. Monitor performance of operators and/or contractors and verify the quality of the data with on-site surveys. Operational plans - stipulating deviations with respect to route schedules and fare management - that need to be implemented in the case of a special event.

At first glance, the measurement tools currently used by the MyCiti system may appear to be relatively all-inclusive. However this is not the case. Not only is the monitoring done by a modular- and possibly inconsistent performance measurement approach, but also insufficient attention is given to the monitoring of the IRT system's inherent technology-related aspects. Moreover, with their current disintegrated approach, it is difficult or impossible to attain an idea of the overall system health. As a result, the sustainability of the MyCiti system is at question.

VI. DISCUSSION

The status quo on the technology deployments presented herein elucidates: 1) the shortcomings- and the unsustainable nature of the current modular measurement systems and 2) the complexities associated with the sustainable implementation of the ITS applications.

VII. RESULTS

By considering the local ITS industry and the larger performance measurement field, Table 2 has been established. Although Table 2 is still in the early stages of development, the fundamental aspects of performance

TABLE 1: MyCiti's Performance Measurement Toolbox

	Performance Area	Name	Type	Age
1	Base System Planning	Divia	Computer Software	2012
2	Vehicle Tracking	Lio		
3	Schedule Adherence			
4	Operational Performance	BI Analysis	Business Objects	2013
5	Financial Performance	AFC Financial Analysis	Microsoft Excel	
6	Driver Behavior Risk Management	Drivecam	Equipment/ Web-based	2011
7	Asset/ Maintenance/ SLA Management	Forcelink	Web-based & Mobile Interface	
8	Public Relations	Marketing	Service Promotion (Media)	2010
9	Customer Care	Customer Interaction	Service Provision (Social Media)	
10	Quality Assurance	Performance & Quality Monitoring	Performance Evaluation & Quality Assurance Checks	2010
11	Event Management	Special Events	Operational Deviation	

measurement and the primary features of ITS projects are addressed. A Sustainable Balanced Scorecard (SBSc) as the reference model for the performance framework has been adopted. This SBSc builds on the done work in [5].

The SBSc is to act as a management tool for assessing the effective development and maintenance of ITS applications as well as for evaluating the feasibility of continuous investments in transportation technology. Horizontally the SBSc is subdivided into five sustainability perspectives. These are: 1) learning and growth, 2) internal processes, 3) financial management, 4) society and 5) environment. Vertically the SBSc is subdivided into performance environments (level 1) and performance areas (level 2). Refer to Table 2.

VIII. CONCLUSIONS

Through the promotion of ITS applications, SA is embracing a technology-driven setting. Several initiatives with regard to the implementation of information systems within the nation's transportation environment support this statement. The CoCT's TMC is an example of this. Since the realization of this transit- and traffic operations' facility, extensive deployment of technology and the relating supporting ITS devices have been implemented.

Moreover, the South African transportation industry at large seems to be aware of the need for measuring performance. Numerous attempts at measurement systems

(although none have yet achieved full functionality) serve as testimony of this fact. SANRAL has developed a KPI approach towards the performance measurement of FMS, the *WhereIsMyTransport* application provides an operational environment toolset to the transportation operator and the MyCiti system has pursued a performance measurement toolbox approach.

At this stage, however, the emphasis on a consistent and holistic performance measurement approach still lacks. Furthermore, insufficient attention is given to the measuring and monitoring of the technology-related aspects of the ITS applications. An all-inclusive and easy-to-understand framework that can stimulate the achievement of the ultimate ITS performance management regime is needed. It is believed that the SBSc developed herein serves as the foundation for such a regime and hence also fosters the achievement of sustainable transportation.

IX. FUTURE RESEARCH

With the completion of the research project on which this paper is based, the SBSc presented in Table 2 will be further developed by identifying each performance area's representative performance measure(s) as well as each performance measure's relating KPI and standard. With the aid of Multiple Criteria Decision Analysis (MCDA) principles, the overall system health will then be determined and portrayed on a consolidated performance dashboard.

ACKNOWLEDGMENT

It is an honor for me to thank Chantal Greenwood from the CoCT's IRT department: TCT (Transportation for Cape Town). She has provided me with very insightful and important data that has enabled me to attain a profound grasp on IRT systems. I am very grateful for her help and efforts.

REFERENCES

- [1] Artley, W., Ellison, D. J. and Kennedy, B., 2001, 'Establishing and Maintaining a Performance-Based Management Program', The Performance-Based Management Handbook, Vol. 1, September 2001, Performance-Based Management Special Interest Group (PBM SIG).
- [2] CoCT, 2009, 'ITP for the CoCT: 2006 - 2011', May 2009.
- [3] Eboli, L. and Mazzulla, G., 2012, 'Performance indicators for an objective measure of public transportation service quality', European Transportation, Issue 51, August 2012, The Institute for the Study of Transportation within the European Economic Integration (ISTIEE).
- [4] Ezell, S., 2010, 'Explaining International IT application Leadership: ITS', The Information Technology and Innovation Foundation (ITIF), January 2010.
- [5] Rabbani, A., Zamani, M., Yazdani-Chamzini, A. and Zavadskas, E. K., 2014, 'Proposing a new Integrated Model based on Sustainability BSC and MCDM Approaches by using Linguistic Variables for the Performance Evaluation of Oil Producing Companies', Expert Systems with Applications, Vol. 41, Issue 16, November 15, 2014, pp. 7316 - 7327, Elsevier Ltd.
- [6] Republic of SA: Department of Transportation (DoT), 2007, 'ITP: Minimum requirements in terms of the NLTA', National Land Transportation Transition Act, No. 22 of 2000, Government Gazette, November 30, 2007.
- [7] SANRAL, 2011, 'Procurement of a National ITS and Integrated Supporting Systems Software and the deployment thereof in Gauteng, Kwazulu-Natal and the Western Cape', Contract Performance Measurement, Vol. 2, Book 6, January 2011, Techso (Pty) Ltd.

TABLE 2: Sustainable Balanced Scorecard

Sustainable Balanced Scorecard											
Learning & Growth	GOVERNANCE, PLANNING AND DECISION MAKING										
	1	<i>Regulation Management</i>				<i>Service-Operation Coordination</i>			<i>Research & Development</i>		
	2	Safety & Security Enforcement	Contract Performance Management	Legal Operational Requirements	Policy Planning	Special Event Planning	Institutional Cooperation	Interoperable Operations	Continuous Improvement	Record Management	Data Management
Internal Processes	SERVICE-PROCESS EXCELLENCE										
	1	<i>Operational Performance</i>			<i>Technology Performance</i>		<i>Emergency & Response Performance</i>		<i>Asset Maintenance & Management</i>		
	2	Service Reliability	Service Quality	Operating Efficiency	Technology Reliability	Technology Quality	Incident Management		Asset Maintenance	Asset Management	
Financial Management	ECONOMIC PROSPERITY										
	1	<i>Economic Contribution</i>	<i>Economic Sustainability</i>			<i>Economic Vitality</i>			<i>Capital Investment Management</i>		
	2	Transportation GDP	System Preservation		System Adaptability	Economic Growth	Economic Development	Economic Impression	Capital Improvements	Investment Management	
Society	SOCIAL EQUITY										
	1	<i>Public Relations</i>			<i>Human Resource Management</i>			<i>Sustainable Communities</i>		<i>Facility Management</i>	
	2	Awareness Service	Customer Care	Employee Performance Management			Sustainable Living		Servicescape		
Environment	ENVIRONMENTAL STEWARDSHIP										
	1	<i>Environmental Conservation</i>						<i>Resource Conservation</i>			
	2	Land-use Management		Climate Change Management		Noise Management		Waste Management		Energy Management	