Multimodal Biometric Surveillance using a Kinect Sensor

Ross Savage[1], Nathan Clarke[1,2] and Fudong, Li[1,]

[1]Centre for Security, Communications and Network Research (CSCAN), School of Computing & Mathematics, Plymouth University, Plymouth, PL4 8AA, United Kingdom
info@cscan.org
[2]School of Computer and Information Science, Edith Cowan University, Perth, Western Australia

ABSTRACT

Surveillance technologies are becoming increasingly important in the fight against crime and terrorism. More recently, technologies such as biometrics are being applied to provide an automated approach to the verification of individuals – removing the need for human operators. Unfortunately, current approaches suffer from a number of issues such as expensive biometric capture devices and reliance upon uni-modal systems. This paper presents a feasibility study into the use of an off-the-shelf commercial device (Microsoft Kinect) as a multimodal biometric sensor. An experimental study is undertaken, combining two biometrics modalities (face and gait recognition) and a soft biometric (skeleton measurements). Based upon 20 participants, the study shows an Equal Error Rate (EER) of 12.3% can be achieved – outperforming the results the single uni-modal approaches were able to obtain.

Keywords: biometrics, multi-modal, surveillance, security, skeleton

1 INTRODUCTION

In the modern society, surveillance has been utilised as a powerful monitoring tool to identify individuals and detect unauthorised activities in many of applications, such as border control in an airport, door access at a building entrance and Closed-circuit television (CCTV). Since the 9/11 terrorist attack, a world-wide demand for surveillance has increased dramatically both in quantities and qualities. For instance, at least 1.85 million CCTV systems have already been deployed within the UK for the purpose of monitoring public security on a daily basis (The Guardian, 2011). Furthermore, systems have been equipped with high-resolution cameras to ensure the caption of clear images and videos. This provides a firm foundation that enables biometric recognition techniques to be deployed on top of existing surveillance systems.

Biometric recognition or biometrics is an automatic process to uniquely identify humans based upon one or more physiological (e.g. face) and/or behavioural (e.g. gait) characteristics (Prabhakar *et al*, 2003). Therefore, biometric surveillance is a system that utilises human biometric traits to automatically identify individuals within a monitored area. For instance, a person can be quickly identified in a large crowd by utilising a face-recognition enabled surveillance system while they walk pass the camera (Mashable, 2012). In additional to existing techniques that are utilised by the traditional biometric system, biometric surveillance can also employ a number of soft biometric methods (e.g. the colour of hair and the height of a body) to narrow down the population of suspects.

Indeed, a number of biometric techniques have already been researched and utilised for the purpose of surveillance. In the US, the Department of Homeland Security invested heavily in facial recognition for identifying known terrorists in airport and bus stations through their surveillance systems (USA today, 2007); this can also be utilised by the police to identify suspects on the street via their CCTV cameras. In the UK, a British e-passport that features an electronic chip containing biometric data (e.g. fingerprint) of the passport holder can be utilised to pass through border controls without interacting with a person using e-passport gates at an airport such as Heathrow, Luton, or Gatwick (Directgov, 2011). That said, the majority of existing surveillance systems rely upon a single biometric technique for the identification process, raising several issues such as performance and circumvention (Vine, 2012; Gulf News, 2012); resulting in a biometric surveillance system having legitimate users rejected and/or imposter accepted. It is arguable that utilising multiple biometric techniques in a single surveillance system can reduce the aforementioned issues. A secondary issue with respect to current surveillance systems is the expensive nature of the biometric sensors (e.g. cameras), particularly with multimodal systems. As a result, this paper conducted a feasibility study into a multi-modal biometric surveillance that employs three biometric techniques (i.e. facial, gait and skeletal (a soft biometric) by using a Microsoft Kinect© sensor.

This paper presents an overview of biometric surveillance, demonstrating the need for a robust biometric surveillance system before proceeding to describe the prior work in the area. In section 3, a series of experiments were conducted to examine the feasibility of utilising the facial, gait and skeletal information to discriminate individuals within a surveillance environment, both individually and as a multimodal system. The paper concludes with a discussion on the impact of the experimental findings.

2 BACKGROUND OF BIOMETRIC SURVEILLANCE

Terrorism has been the single largest factor that has driven the need for biometric surveillance. Many law enforcement agencies and surveillance product manufactures have heavily invested in the area of biometric surveillance systems and a lot of these systems have been deployed in high sensitive security areas (e.g. airport) (USA today, 2007; Directgov, 2011). Indeed, biometric surveillance systems have also found the way into council-based CCTV operations.

Based upon the biometric characteristics, these biometric surveillance systems can be categorised into either physiological or behavioural based, both of which will be discussed in the following sections.

2.1 Physiological Biometric Surveillance

The physiological biometric surveillance systems utilise the characteristics of a human body part to identify individuals. In general, physiological biometric characteristics are resistant to various factors which may affect their performance. For instance, people's fingerprints will not be affected by their age, mood, body fitness or the weather conditions. Moreover, an individual's physiological biometric characteristics contain high levels of discriminatory information. A number of physiological biometric techniques that can be utilised for the purpose of surveillance are described as below.

Facial recognition is a technique to identify people through their facial characteristics, such as the distance between the eyes, width of the nose, the shape of cheekbones and the depth of eye sockets (Bledsoe, 1966; Goldstein *et al*, 1971). The technique is user friendly because a face photo can be taken from a distance without any user interaction. As a result, the identification process can be performed secretly or covertly without the user's knowledge;

however system performance can be affected when a poor quality photo is taken. The approach can be easily integrated into a video surveillance system and many CCTV cameras now have extremely high resolutions to facilitate this – up to 29 megapixels (BBC News, 2012). To date, many facial recognition surveillance systems that have the ability of identifying a particular person amongst a large crowd have been designed and deployed for the purpose of public safety, such as AxxonSoft (2011), OmniPerception (2012), RT (2012).

Iris recognition identifies users by examining their iris. The iris is the coloured muscle surrounding the human eye pupil and it is highly unique to each individual person (Daugman, 1993). In order to obtain an iris image, one of the following types of camera is used: near infrared (NIR), high-resolution visual light and telescope-type. Moreover, the initial cost for the equipment can be very expensive, especially for long range cameras. As a result, iris recognition surveillance systems have only been implemented for applications requiring high security. For example, the Iris Recognition Immigration System (IRIS) that is an iris recognition surveillance system is currently being deployed to identify passengers by the UK Border Agency in several airports, such as Heathrow, Gatwick and Birmingham (UKBA, 2011). Whilst research in being undertaken for covert acquisition of iris images, commercial systems still rely upon an intrusive sample being provided from a willing individual.

In addition, other physiological biometrics such as ear recognition (Abaza et al, 2010) and facial thermography (Prokoski *et al*, 1992) have also been suggested for the purpose of surveillance.

2.2 Behavioural Biometric Surveillance

Behavioural biometric surveillance identifies a person based upon their unique behaviour, such as the way they walk. Human behaviour can change over time due a variety of reasons; aging, fitness, social networking environments and weather conditions are all potential examples. As a result, the discriminatory characteristics also tend to change, affecting the performance of any behavioural biometric surveillance system. Typically the performance obtained in behavioural-based systems is weaker due to their less stable feature set.

Gait recognition employs a sequence of human limb movement to identify an individual (Xu et al, 2006). Gait motion can be obtained using a camera or video recorder (images can be extracted from the video) in the form of a sequence of pictures. As the gait images are obtained from a distance without any user physical contact, a gait recognition based system is a non-intrusive approach. However, a person's gait can and does change over a long period of time due to their age, body weight or fitness. In addition, gait can be influenced by other factors, such as the weather, footwear, ground conditions and personal emotions. As a result, the performance of a gait recognition based system can vary. Gait recognition applications could potentially be used for video based intelligent surveillance systems to verify people's identity in the future (Lu and Zhang, 2007; Hossain and Chetty, 2012).

Voice verification is based upon the way how people speak (i.e. voice speed and speaking accent) to identify individuals (Campbell, 1997). Voice verification can operate in three modes: static (word dependent), dynamic (word independent) and pseudo-dynamic. For the purposes of surveillance, only the dynamic-based approach is feasible – as identify verification does not depend upon a predefined or prescribed spoken phrase. Unfortunately, in terms of performance, this approach given its increased complexity tends to perform poorly in comparison to the other static-based approaches.

Apart from the aforementioned techniques, other behavioural-based biometric methods can also be utilised in a surveillance system, such as human shadows (Iwashita and Stoica, 2009; Iwashita *et al*, 2012), human body parts (Denman *et al*, 2009), and the colour of eye, skin and hair (Dantcheva *et al*, 2010). However, these approaches can largely be categorised as soft biometrics – discriminative information that can help reduce a population search but not sufficient distinct to uniquely identify an individual.

2.3 Summary on Biometric Surveillance

Biometric surveillance is able to automatically identify an individual amongst a large crowd and law enforcement officers can quickly find suspects in public areas (e.g. airports and train stations). However, majority of the existing biometric surveillance only employ a single human characteristic (e.g. face, iris, gait), raising several issues such as performance, circumvention and single point of failure (Vine, 2012; Gulf News, 2012; The H Security, 2011; Venugopalan and Savvides, 2011). Although a number of studies have utilised several modalities together (Denman *et al*, 2009; Dantcheva *et al*, 2010), they could only be utilised to narrow down suspects within a large crowd but not to accurately identify individual. Therefore, a biometric surveillance system that can provide robust security is needed. Unfortunately, the cost of biometric sensor technology, particularly multimodal hardware has been traditional prohibitively expensive. This study looked to examine the capability that can be achieved through utilising off-the-shelf hardware and software.

3 EXPERIMENTAL METHODOLOGY

An investigation of possible biometric sensor technologies resulted in the Microsoft Kinect© being identified as a device capable of multimodal capture of face, gait and skeleton signals. The Kinect has two key advantages, the ability to capture multimodal signals within a single piece of hardware and was cost effective compared to the remaining options. In order to undertake the experiment in order to collect and transform the image and measurement signals into various biometric templates, a data capture application was developed in the C# language supported via the Kinect© Software Development Kit (SDK) (Microsoft, 2012). As shown in Figure 1, the data collection application has a graphical control panel that permits an administrator to complete a number of tasks, such as participant enrolment, the folder creation for the individual data captures of face, gait, and skeletal measurements, and the collection of multiple biometric samples.

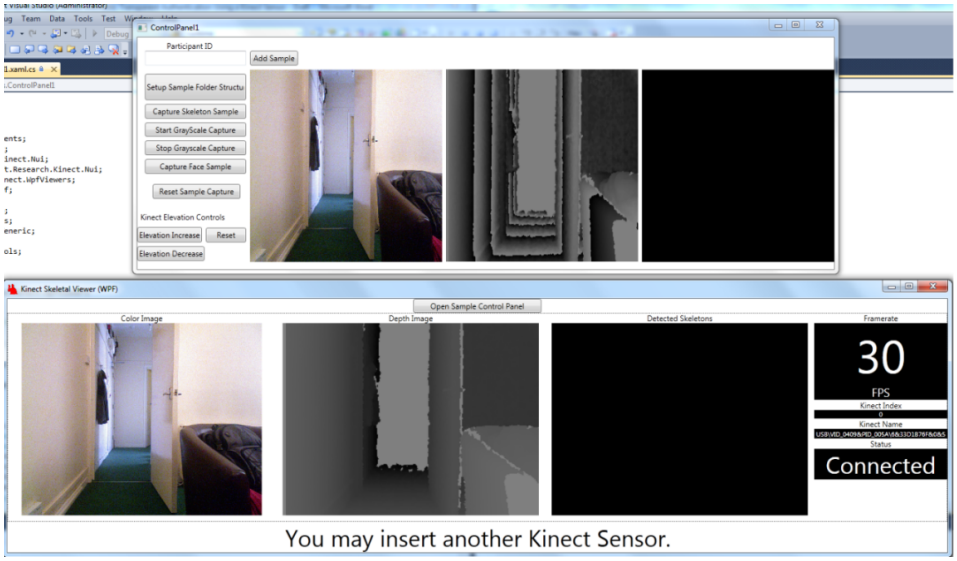

Figure 1: Screenshot of the Data Capture Application

The experiment involved 20 participants. Each participant was asked to walk 3 meters with the camera in a head-on orientation. Whilst gait recognition is typically captured perpendicular to the direction of motion, the requirement to capture Skelton and face samples simultaneous necessitated a heads-on capture. Both skeleton and gait measurements were taken whilst the user was in motion. When the user stopped at the end of 3 meters, a face-based sample was captured. Each participant was required to perform the process a total of 15 times, allowing sufficient experiment data to be collected for both enrolment and verification.

In order to process and analyse the biometric samples, MatLab was utilised. MatLab is an industry-accepted modelling and simulation environment that enabled the manipulation of processing the sample data.

The processing and evaluation of the biometric modalities conforms to standard biometric testing procedures. 4 of the samples were used in the enrolment process, with the reaming 11 samples utilised in the verification process. Each user was given the opportunity of acting as an authorised user, with the remaining users acting as impostors. It is worth highlighting that the experiment was devised to operate in verification rather than identification mode – whilst true surveillance systems operate in identification mode, the experiment was limited to the easier mode of identity verification in the first instance. In reality this would mean such a system has a more specific role in acting as a surveillance system for known individuals. For example, this system could be deployed a passport control, with the sensor capturing data as the individual walks up to the passport counter. The passport would provide the "who I claim to be" credential from which the biometric samples could then be evaluated against.

Three classifiers were used for evaluating the performance of individual techniques: fisher faces algorithm, neural networks and a simple standard deviation method for facial, gait and skeletal respectively. The first two approaches were extract from algorithms developed by AdvancedSourceCode (2013). Whilst these algorithms were not free, they were incredible cheap and are made available without any licensing restrictions or costs – making them effectively free in comparison to commercial-grade classifiers. The multi-modal technique, utilised a fusion based classier that has the ability to apply different weighting for individual biometric modalities. An evaluation of each of the individual modalities and of the fusion approach was undertaken in order to provide a comparison on the performance.

4 RESULTS

The overall system performance for facial recognition using fisher's algorithm is shown in figure 2. The equal error rate is 13.2%. The overall system performance for gait recognition using a feature based algorithm is shown in figure 3. The equal error rate is 42.7%. Whilst the error rate for gait recognition is poor, it must be noted that the mode of operation- heads-on, rather than perpendicular to, would have a significant impact on the performance.

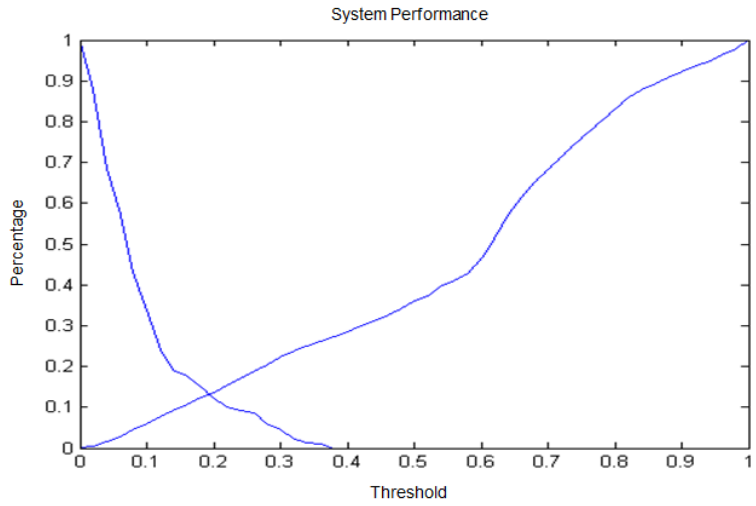

Figure 2: Face Recognition Overall System Performance

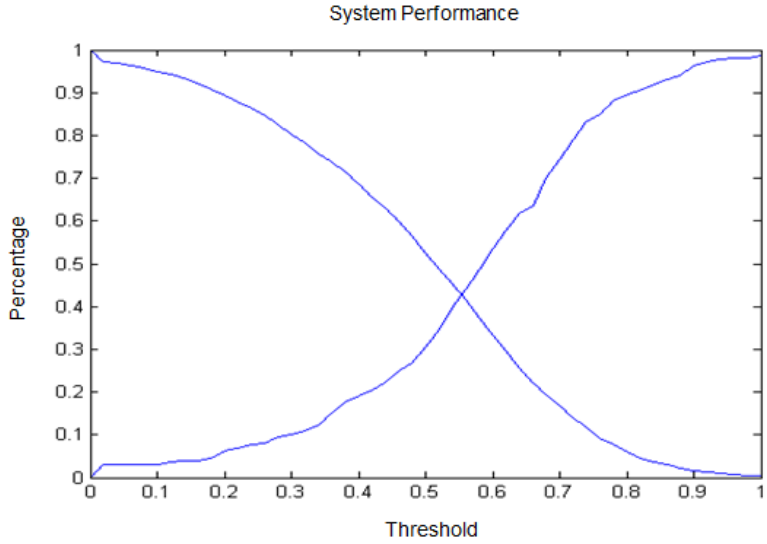

Figure 3: Gait Recognition Overall System Performance

The receiver operating characteristic (ROC) graph (shown in figure 4) shows the 5 most discriminatory standard deviation values. From the curves the optimum standard deviation is 3. The overall system performance for skeletal recognition using a standard deviation based algorithm set to 3 is shown in figure 5. The equal error rate is 38.4%.

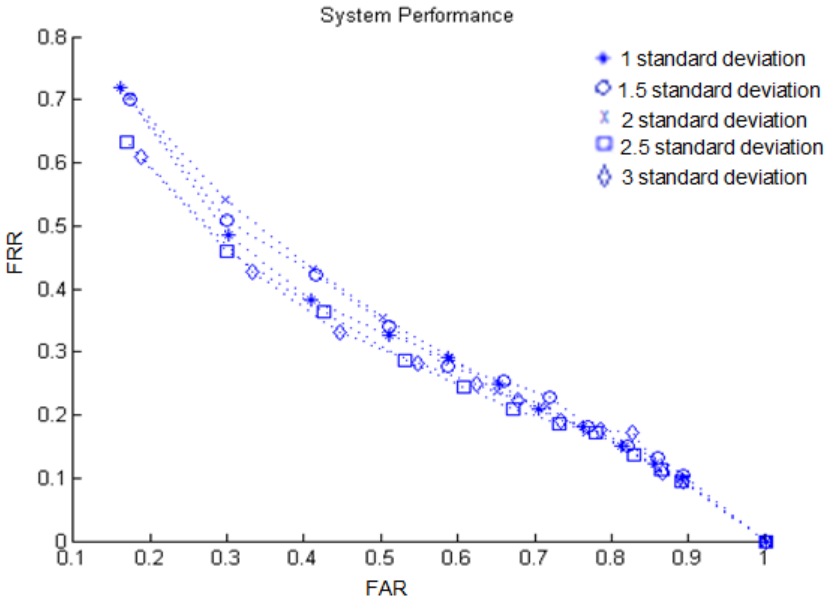

Figure 4: ROC Curve Showing 5 Standard Deviations for Skeletal Measurement Recognition

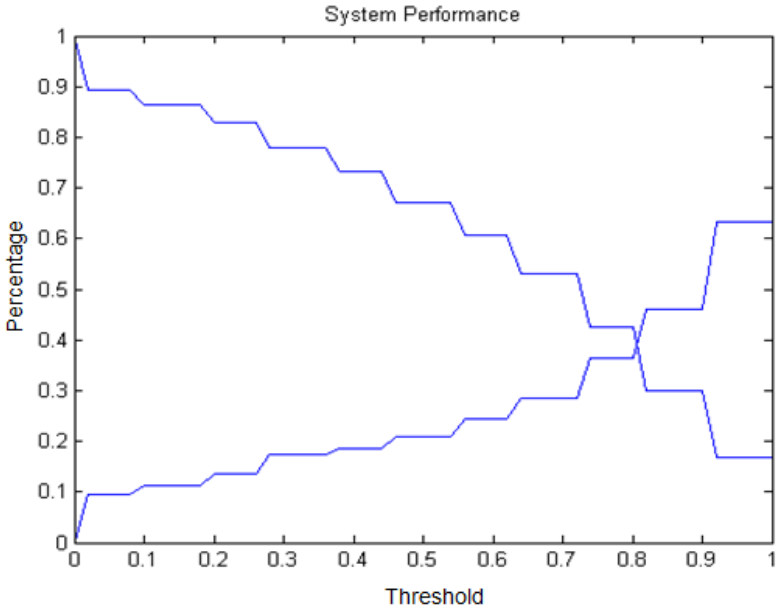

Figure 5: Overall Skeletal Measurement Performance for Three Standard Deviations

Skelton measurements would not be discriminative to be incorporated as a biometric modality; however, the results have shown the value the approach has as a soft biometric in providing additional discriminative information. In this experiment, albeit with limited participants, it outperformed the gait-based modality.

The fusion of these three approaches results in a set of results that indicates that multimodal-based approaches have a degree of usefulness. As illustrated in Table 1, the best EER of a single modality was 13.2%, with the fusion-based approach achieving 12.3%. Unsurprisingly, the poor performance of gait recognition has resulted in that modality adding little discriminative power to the approach.

Weighting	**System EER**
Equal Weighting	15.5%

Face Weighting – 0.8 Gait Weighting – 0.1 Skeletal Weighting – 0.1	12.6%
Face Weighting – 0.5 Gait Weighting – 0 Skeletal Weighting – 0.5	13.5%
Face Weighting – 0.6 Gait Weighting – 0.1 Skeletal Weighting – 0.3	12.9%
Face Weighting – 0.55 Gait Weighting – 0.15 Skeletal Weighting – 0.3	12.7%
Face Weighting – 0.8 Gait Weight – 0 Skeletal Weight – 0.2	12.3%
Face Weighting – 0.7 Gait Weighting – 0 Skeletal Weighting – 0.3	12.9%

Table 1: Multimodal Fusion Results

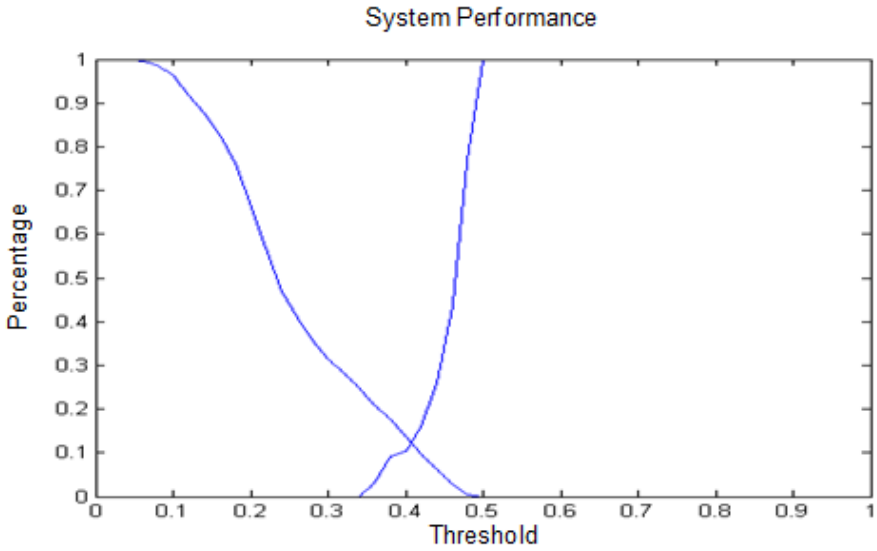

Figure 6: Mutlimodal Fusion Performance Chart

A further examination of individual results does demonstrate a significant variance in the performance individuals are able to achieve. As illustrated in Table 2, individual users can experience EERs as low as 1% but also as high as 31% (ignoring the gait recognition due to its poor performance).

Approach	**Best User System EER**	**Worst User System EER**
Face	1%	31%
Gait	17.9%	60%
Skeletal	25%	52%

Best Multi-Modal Fusion	< 1%	18.3%

Table 2: Multimodal Fusion – Analysis of Individual Performances

5 DISCUSSION

Whilst the overall results do raise a number of concerns over the performance that can be obtained, the results have shown that:

- Off-the-shelf components can be utilised to provide a level of identity verification. Foremost, the use of the Kinect sensor has revealed its usefulness in capture suitable quality samples for use in biometrics
- A good level of facial recognition performance can be obtained from entry-level algorithms
- Whilst supported, heads-on gait recognition is currently unsuitable for use within systems and further research is required examining suitable extraction and classification algorithms
- Skeleton measurements have provided a useful soft-biometric approach
- Multimodal-based approaches will improve the performance that can be achieved via a uni-modal approach.

A further examination of the gait recognition samples did raise an issue that will have had an effect upon the performance results. The silhouettes produced show that the capture software failed to capture two complete walking cycles. A review of the methodology found, that whilst 3 meters was sufficient for the participant to walk two cycles, the time taken for the software to operate and the field of vision provide by the Kinect device resulted in part of the gait cycle not being captured. This problem was a result of the rather restricted physical space available to undertake the experiment and it is not expected to be an issue in more normal circumstances. Furthermore, previous research has shown gait recognition can achieve recognition rates of 95% and therefore should not be completed discounted due to these results (Matovski *et al*, 2012).

The improvement in performance using multimodal fusion is inline with another similar paper, which sought to combine face recognition with soft biometrics (demographic information and facial marks) (Jain and Park, 2010). The lowest overall system EER achieved by the proposed multi-modal approach, did better than face by 0.9%. Whilst this isn't a huge reduction it is a rather notable achievement as it achieved on a consumer device capable of being used as multi-modal system out of the box; rather than having to adapt it from several uni-modal capture devices. Whilst not the lowest EER produced, an overall system EER of 12.7% seen in table 2 shows all three biometric techniques can be utilised to reduce the system EER.

6 Conclusion and Future WORK

The paper has presented the results of utilising an off-the-shelf consumer device to provide biometric-based surveillance. The Kinect has proven to be a robust and effective capture device, capturing images and calculating measurements with sufficient quality to be incorporated reliably within biometric algorithms.

The recognition performance of using gait recognition in heads-on rather than perpendicular mode has resulted in a significant performance drop. Whilst some problems did arise due to the capture software and experimental methodology, the classification algorithms clearly need further research in order to adapt to the different and more confined set of features that exist.

The results have also clearly shown the positive effects skeleton measurements can have. Indeed, the application as a uni-modal approach provided a reasonable set of results – particularly for some individuals. Whilst it is expected this performance would worsen with larger population samples, its use within a multimodal fusion system adds to the overall performance that can be obtained.

Acknowledgements. The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement 284863 (FP7 SEC GERYON).

REFERENCES

Abaza, A., Hebert, C. and Harrison, M.A.F. (2010) "Fast learning ear detection for real-time surveillance," Biometrics: Theory Applications and Systems (BTAS), 2010 Fourth IEEE International Conference on, vol., no., pp.1-6, 27-29 Sept. 2010 doi: 10.1109/BTAS.2010.5634486

Advancedsourcecode (2013) "Gait Recognition System - Matlab source code", http://www.advancedsourcecode.com/gaitrecognition.asp, 15 January 2013

AxxonSoft (2011), "Face Recognition", http://www.axxonsoft.com/integrated_security_solutions/face_recognition/index.php?phrase_id=3032106, 09 July 2012

BBC News (2012), "High-def CCTV cameras risk backlash, warns UK watchdog" http://www.bbc.co.uk/news/technology-19812385, 31 January 2013

Bledsoe, W. W. (1966) "Man-Machine Facial Recognition: Report on a Large-Scale Experiment", Technical Report PRI 22, Panoramic Research, Inc., Palo Alto, California.

Campbell, J. P. (1997) "Speaker Recognition: A Tutorial", Proceedings of the IEEE, volume 85, No. 9, September 1997, pp.1437-1462.

Dantcheva, A., Dugelay, J. L. and Elia, P. (2010) "Person recognition using a bag of facial soft biometrics (BoFSB)," *Multimedia Signal Processing (MMSP), 2010 IEEE International Workshop on* , vol., no., pp.511-516, 4-6 Oct. 2010, doi: 10.1109/MMSP.2010.5662074

Denman, S., Fookes, C., Bialkowski, A. and Sridharan, S. (2009) "Soft-Biometrics: Unconstrained Authentication in a Surveillance Environment," *Digital Image Computing: Techniques and Applications, 2009. DICTA '09.* , vol., no., pp.196-203, 1-3 Dec. 2009 doi: 10.1109/DICTA.2009.38

Daugman, J. (1993) "High confidence visual recognition of persons by a test of statistical independence", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15(11), pp. 1148-1161.

Directgov (2011) "*ePassport gates now available at every major UK airport : Directgov – Newsroom*", http://webarchive.nationalarchives.gov.uk/20121015000000/www.direct.gov.uk/en/Nl1/Newsroom/DG_198686, 20 Octorber 2012

Goldstein, A.J., Harmon, L.D., and Lesk, A.B. (1971) "Identification of Human Faces", proceeding of IEEE, Vol, 59, No. 5, pp.748-760, May 1971.

Gulf News (2012) "New technology puts end to passport fraud", http://gulfnews.com/news/gulf/uae/government/new-technology-puts-end-to-passport-fraud-1.1021716, 21 June 2012

Hossain, E. and Chetty, G. (2012) "Person identification in surveillance video using gait biometric cues," *Fuzzy Systems and Knowledge Discovery (FSKD), 2012 9th International Conference on* , vol., no., pp.1877-1881, 29-31 May 2012

Iwashita, Y. and Stoica, A. (2009) "Gait Recognition Using Shadow Analysis," *Bio-inspired Learning and Intelligent Systems for Security, 2009. BLISS '09. Symposium on*, vol., no., pp.26-31, 20-21 Aug. 2009 doi: 10.1109/BLISS.2009.28

Iwashita, Y., Stoica, A. and Kurazume, R.(2012) "Finding People by their Shadows: Aerial Surveillance Using Body Biometrics Extracted from Ground Video," *Emerging Security Technologies (EST), 2012 Third International Conference on* , vol., no., pp.43-48, 5-7 Sept. 2012, doi: 10.1109/EST.2012.41

Jain, A. and Park, U. (2010) "Face Matching and Retrieval Using Soft Biometrics". IEEE Transactions on Information Forensics and Security, September, 5(3), pp. 406-415.

Lu, J. and Zhang, E. (2007) "Gait recognition for human identification based on ICA and fuzzy SVM through multiple views fusion", Pattern Recognition Letters, Vol. 28, pp. 2401–2411, 2007.

Mashable (2012) "Surveillance System Can Recognize a Face From 36 Million Others in One Second [VIDEO]", http://mashable.com/2012/03/23/hitachi-face-recognition/, 28 January 2013

Matovski, D., Nixon, M., Mahmoodi, S. and Carter, J. (2012) "The Effect of Time on Gait Recognition Performance", IEEE Transactions on Information Forensics and Security, April, 7(2), pp. 543-552.

Microsoft (2012) "Kinect for Windows", http://www.microsoft.com/en-us/kinectforwindows/develop/developer downloads.aspx, 30 January 2013

OmniPerception (2012) "CheckPoint.S™ Real-Time Facial surveillance", http://www.omniperception.com/products/checkpoints-facial-surveillance/, 20 December 2012

Prabhakar, S., Pankanti, S. and Jain, A.K. (2003) "Biometric recognition: security and privacy concerns", *Security & Privacy, IEEE*, vol.1, no.2, pp. 33-42, Mar-Apr 2003, doi: 10.1109/MSECP.2003.1193209

Prokoski, F.J., Riedel, R.B. and Coffin, J.S. (1992) "Identification of Individuals by Means of Facial Thermography", in Proceedings of the IEEE International Conference on Security Technology, Crime Countermeasures, 1992, pp.120-125

RT (2012) "FBI begins installation of $1 billion face recognition system across America", http://rt.com/usa/news/fbi-recognition-system-ngi-640/, 29 January 2013

The Guardian, (2011) "You're being watched: there's one CCTV camera for every 32 people in UK", http://www.guardian.co.uk/uk/2011/mar/02/cctv-cameras-watching-surveillance, 10 December 2012

The H Security (2011) "*Android 4.0 face recognition flawed - The H Security: News and Features*", http://h-online.com/-1379290, 25 March 2012

UKBA (2011) "Using the iris recognition immigration system (IRIS)", http://www.ukba.homeoffice.gov.uk/travellingtotheuk/Enteringtheuk/usingiris/, 05 May 2012

USA today (2007) "Face recognition next in terror fight", http://usatoday30.usatoday.com/news/washington/2007-05-10-facial-recognition terrorism_N.htm, 31 January 2013

Venugopalan, S. and Savvides, M., 2011. How to Generate Spoofed Irises From an Iris Code Template. *Information Forensics and Security,* June, 6(2), pp. 385-395.

Vine, J (2012) "Inspection of Border Controls at Heathrow Terminal 3", http://icinspector.independent.gov.uk/wp-content/uploads/2012/05/Inspection-of-Border-Control-Operations-at-Terminal-3-Heathrow-Airport.pdf, 20 December 2012.

Xu, D., Yan, S., Tao, D., Zhang, L., Li, X., and Zhang, H. (2006) "Human Gait Recognition With Matrix Representation," *Circuits and Systems for Video Technology, IEEE Transactions on Circuits and Systems for Video Technology*", vol.16, no.7, pp.896-903, July 2006, doi: 10.1109/TCSVT.2006.877418