

Quis-Campi: Extending In The Wild Biometric Recognition to Surveillance Environments

João C. Neves¹, Gil Santos¹, Sílvia Filipe¹, Emanuel Grancho¹, Silvio Barra², Fabio Narducci³, and Hugo Proença¹

¹ Department of Computer Science, IT - Instituto de Telecomunicações
University of Beira Interior, Covilhã, PORTUGAL

² DMI - Dipartimento di Matematica e Informatica
University of Cagliari, Cagliari, ITALY

³ DISTRA-MIT, University of Salerno, Fisciano, Salerno, ITALY

Abstract Efforts in biometrics are being held into extending robust recognition techniques to *in the wild* scenarios. Nonetheless, and despite being a very attractive goal, human identification in the surveillance context remains an open problem. In this paper, we introduce a novel biometric system – *Quis-Campi* – that effectively bridges the gap between surveillance and biometric recognition while having a minimum amount of operational restrictions. We propose a fully automated surveillance system for human recognition purposes, attained by combining human detection and tracking, further enhanced by a PTZ camera that delivers data with enough quality to perform biometric recognition. Along with the system concept, implementation details for both hardware and software modules are provided, as well as preliminary results over a real scenario.

1 Introduction

Biometrics is one of the most active fields in the area of computer vision, which is justified by our societies’ increasing concern about security. Biometric systems significantly rely on the accurate extraction of individuals’ distinctive features, which is conditioned by the acquisition environment and constraints. As such, the most reliable systems are deployed on controlled scenarios and count on subject cooperation. On the other hand, surveillance cameras are widely deployed and can constitute a good source of input for biometric systems. Filling the gap between biometrics and visual surveillance is quite a desirable goal, allowing to produce *automata* capable of recognizing human beings *in the wild*, without their cooperation and, possibly, even without their awareness.

When moving to *in the wild* scenarios the acquisition constraints are substantially lowered and, most of the time, subject cooperation is not even expectable. In order to deal with such challenging conditions, alternatives are sought over three axes [6]: 1) improve the existing algorithms so they can handle more degraded data; 2) resort to multi-modal biometric systems so that the usage of multiple traits can compensate for their lack of “quality”; 3) explore new biometric traits that could better cope with this new reality. Despite the recent efforts, no system yet exists capable of dealing effectively with all the issues introduced by *in the wild* biometrics, and even those systems able to cope with less constrained conditions (e.g. the Iris On The Move project [11])

still lack an ideal level of user abstraction. Most of existing surveillance systems are focused on activity recognition (e.g. W^4 project [5]), and not that many of them are prepared to handle surveillance scenarios by a watchlist approach (e.g. Kamgar-Parsi *et al.* [8]). In this paper, we present a novel biometric recognition system, designed to work covertly in a non-habituated and non-attended fashion, over non-standard environments. Our main goal is to conceive a system that links together both biometrics and visual surveillance, being able to conduct biometric recognition over typical surveillance scenarios, with the minimum possible amount of operational restrictions.

The remainder of this paper is organized as follows: in Sect. 2 we detail the three layers of the recognition system, its operation premises and devised modules; in Sect. 3 we present the exploited techniques for each module, along with preliminary results of our system over a real surveillance scenario and, finally, Sect. 4 states some final considerations.

2 The QUIS-CAMPI System

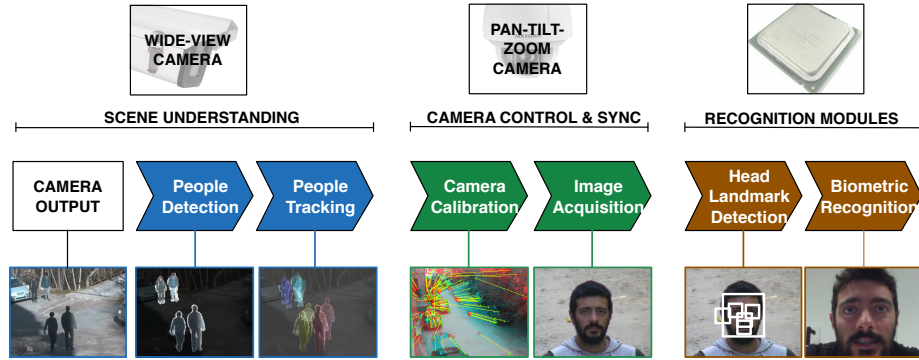


Figure 1. Working diagram of the proposed system, and the three-layer architecture: scene understanding, camera control/synchronization and recognition modules.

The optimal recognition system would operate in any environment, thus minimizing the amount of operational restrictions. Aiming at bridging biometrics with the visual surveillance, we have developed our system in a typical surveillance scenario – a parking lot (Fig. 3(a)) – particularly harsh for recognition purposes: 1) it is a non-standard environment with irregular lighting that changes during the day and accordingly to weather conditions, reflections, etc.; 2) complex background regions and the varying resolution of humans poses increasing challenges for both detection and recognition phases; 3) subjects can come from any direction, and they are rarely facing the camera which is typically placed on an upper position.

To develop such a system we combine a PTZ with a typical surveillance camera in a master-slave configuration. We believe that this architecture is able to provide enough

quality for biometric recognition at-a-distances (15 to 35 meters), since the mechanical properties of the PTZ camera allow to acquire high-resolution imagery of arbitrary locations in the scene. The advantages of using PTZ cameras for biometric recognition are further evidenced by Fig. 2, where the resolution differences between using a wide-view camera and using a PTZ device are evident.

The proposed system is thus devised over three main layers (Fig. 1): scene understanding, camera control/synchronization, and recognition modules. Scene understanding refers to the detection and tracking of human beings. This phase should be supported by the wide-view camera so that it provides head location of persons in the scene, allowing the PTZ camera to zoom-in on those regions. Following the PTZ image acquisition, the recognition modules are responsible to infer the identify of the subject.

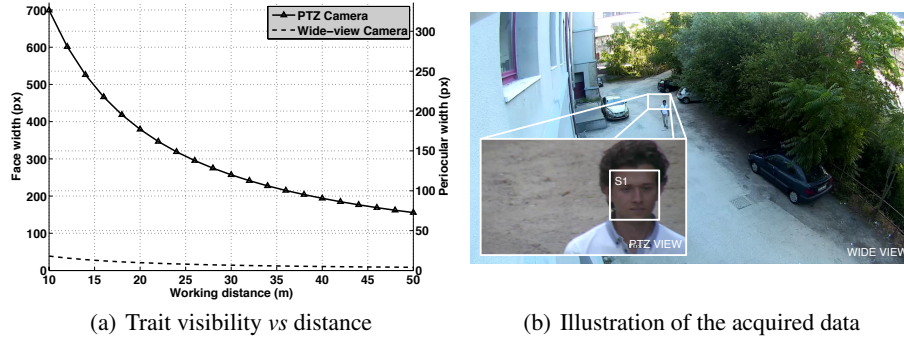


Figure 2. Visible face and periocular width, in pixels, as function of the system’s working distance (a), and illustration of the acquired data for both cameras (b).

2.1 Scene Understanding

The scene understanding layer has two main modules: people detection and tracking. The first module locates persons as they enter the scene and tracks them until they are no longer visible, taking as input the video feed from the wide-view camera, and has three main steps: background subtraction, upper-body detection and tracking – Fig. 3.

2.2 Camera Control and Synchronization

Considering that the wide-view and the PTZ can be disposed arbitrarily in the scene, a calibration algorithm is required to relate the image coordinates of both devices. However, due to the lack of depth information, this problem is ill-defined and thus several approximations have been proposed to alleviate the inaccuracies of 2D-based methods. With a view to determine a precise mapping the devices, different solutions have been proposed to infer 3D information from the scene. In our system, we rely on [19] where the subjects height is inferred and used as an ancillary measure to define a precise mapping between the cameras.

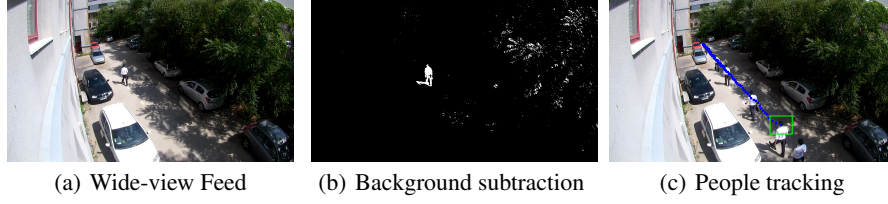


Figure 3. Illustration of the preliminary results obtained by the people detection and tracking module: a) sample image acquired with the wide-view camera; b) foreground regions attained by background subtraction; c) people tracking module results.

Additionally, it is necessary to plan, in real-time, the sequence of PTZ observations when multiple subjects are in the scene. Despite a random walk could be adopted, this strategy would lead to failures in the observation of some targets as the number of subjects increases. For this purpose, we rely on [20] where an algorithm for maximizing the observed number of targets has been devised.

2.3 Recognition Modules

After a successful acquisition of a PTZ shot, the recognition module should be supported by a head landmark detection phase. This strategy improves recognition performance since it determines which facial landmarks are visible, and thus decides the weight of each recognition module. Being able to describe which facial traits are visible and where, is far more important than actually getting a close estimation of the head's pose, as we can tell to which extent the trait is reliable or not.

For recognition purposes, the proposed system relies on a multi-modal biometric approach that combines face, iris, periocular, ear shape and gait information.

The face is not only one of the most common and widely used biometric trait, but also one of the most successful applications of image analysis and understanding, with a lot of techniques available [18]. As the “great variability in head rotation and tilt, lighting intensity and angle, facial expression and aging” make face recognition an extremely hard challenge [2], it is mandatory to rely on robust approaches (e.g. [17]). Face recognition algorithms are based either on the global analysis of the whole image, or the relation between facial elements, their location and shape. The main drawbacks are: the 3D structure of the face, which leads to altered appearance accordingly to subject's pose; the occlusion of large portions of non-orthogonal data acquisition; the changes in appearance introduced by facial expressions; and the easiness in disguise. These factors become even more evident *in the wild*, or with uncooperative subjects who try to avoid detection.

The ocular region is one of the most explored in biometrics. Iris in particular is a very popular biometric trait, delivering very high recognition accuracy under controlled environments. Although iris performance as a biometric trait being severely impacted in non-ideal setups, due essentially to its reduced size and moving profile, researchers are putting efforts in overcoming those limitations. The periocular region represents a good trade-off between the whole face and the iris alone, being easy to acquire without user

cooperation, and not requiring a constrained close capturing, being one of the strongest candidates for the purposes of our system.

The shape of the ear can also be used as a biometric trait, as the structure of its cartilage is unique for each individual and its patterns can be imaged on the visible wavelength (VIS) with regular cameras. Despite all ear recognition methods traditionally require some degree of user cooperation, if proper alignment estimation can be established it can be used as biometric trait *in the wild*.

Gait is the only trait that will be imaged from the wide-view cam. Acquiring data about the way a person walks is non-invasive, and can be done at-a-distance. The majority of the gait recognition methods in the literature do not require high-resolution data, so they can run over surveillance camera data.

3 Experimental Results

This section details the exploited techniques for each module, along with preliminary results over the selected surveillance scenario. In our experiments a wide-view camera (Canon VB-H710F) and a PTZ camera (Hikvision DS-2DE5286-AEL) were mounted on the exterior of a building at a first-floor level (approximately 5m above the ground) pointing towards a parking lot.

3.1 People Detection and Tracking

For the background subtraction step, SOBS [10] and Mixture of Gaussians [14] were used. This option was taken after visually inspecting multiple state-of-the-art techniques' performance over test data. Using the output from the background subtraction, we filtered the most consisted regions with human presence by exploiting an upper body detector based on Haar feature-based cascade classifiers [16].

The tracking phase is then initialized, exploiting motion and appearance features. Using the omega-shape (head and shoulder region) as the primary source of key-points, the Kanade-Lucas-Tomasi (KLT) algorithm [13] tracks the initial set of features accordingly to motion and appearance constraints. Since some features may be lost during the process, re-initialization of the features is ensured by the detection phase – Fig. 3(c). The KLT algorithm was preferred since it assumes that a set of discriminant points of the object move with a constant speed and maintain a constant appearance. Based on the set of previous locations provided by the tracking module, a Kalman filter [7] is used to provide a coarse estimation of the future position. We observed that although maintaining their exterior looking while passing through the scene, dynamic lighting and shadow interference perturb persons' appearance. On the contrary, people moving at constant speed provide higher confidence on motion features.

To assess the reliability of the proposed method for tracking, we considered four different scenarios with increasing level of difficulties: *S1*- single person, moving away from the camera, noiseless background subtraction, no significant changes in lighting; *S2*- single person, moving away from the camera at a higher speed, some noise in the background subtraction, no significant changes in lighting; *S3*- single person, walking towards the camera, significant noise in the background subtraction, significant lighting

changes; *S4*- three persons, moving away from the camera, little noise in background subtraction, one of the subjects crosses the path of the other two.

Results are presented in Table 1, using the CLEAR-MOT metrics (Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), True Positive Rate (TPR), False Positive Rate (FPR) and mismatch (MIS)), a standard metric for evaluating multiple target tracking algorithms [9].

Table 1. Tracking performance in our surveillance scenario, when using KLT. Performance metrics are Multiple Object Tracking Accuracy (MOTA), Multiple Object Tracking Precision (MOTP), True Positive Rate (TPR), False Positive Rate (FPR) and mismatch (MIS).

Scenario	MOTA	MOTP	TPR	FPR	MIS
S1	0.940	0.600	0.970	0.030	0
S2	0.800	0.590	0.900	0.100	0
S3	0.745	0.336	0.862	0.138	0
S4	0.589	0.288	0.792	0.202	3

High levels of accuracy (MOTA) and precision (MOTP) were obtained for the first scenario, with a negligible FPR, mostly due to the high quality mask from background subtraction which led to a very precise tracking. Regarding scenarios *S2* and *S3*, all MOTA, FPR and TPR confirm encouraging levels of performance of the tracking algorithm. The significant change in precision observed in *S3* comparatively to *S2* is related to the distance that the subject enters the scene: at long walking distances the number of pixels representing a person is very small, leading to a failure of the upper body detector. In the most challenging scenario (*S4*) the FPR increases, along with some mismatches related essentially to the path crossing between persons. Nonetheless, we can assert that the tracking method achieves good level of performances. Although the omega shape at long distance being hard to detect, the whole body shape should be a better alternative, and once a tracker detects the shape of a person, the head will appear on the top of the selected area.

3.2 Biometric Recognition

To have a preliminary assessment of the recognition performance of our system, 20 participants were imaged between distances 15 to 35 meters. These working distances ensure regions with widths between 500 px and 200 px in the face, and approximately 220 px to 100 px for the periocular region. Facial region was determined using a cascade object detector based on Viola and Jones algorithm [16], and facial features encoded using the Principal Component Analysis (PCA) approach [15]. Prior to encoding the periocular features, a second Region of Interest (ROI) containing the periocular region was defined still using a Viola and Jones based cascade object detector, trained for the detection of the right eye using Haar features to encode the details [3]. Upon that region, five different descriptors were extracted, based on the works of Park *et al.* [12] and Bharadwaj *et al.* [1]: Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), Scale-Invariant Feature Transform (SIFT), Uniform Local Binary Patterns (ULBP) and

GIST. The HOG, LBP and ULBP descriptors deliver a distribution-based analysis, and were computed over 35 non-overlapping patches of the periocular ROI, evenly distributed on a 7×5 grid. Each descriptor was computed sequentially, forming a global 1-D array storing both shape and texture information.

Finally, two score-level fusion were also stressed: one combining the scores from the individual periocular recognition methods; and a second one combining them with the PCA results. Score fusion was achieved training a Neural-Network (NN) with two hidden layers using back-propagation. NN based methods are widely applied in classification problems, for their learning abilities and good generalization capabilities. The architecture of the used NN consisted on a first hidden layer with the number of neurons equalling the number of scores to be fused, and a second hidden layer of three neurons. The final (output) layer had one neuron, since we were dealing with a binary classification problem. The NNs were trained with a smaller partition of the data, not included on the test phase.

Three metrics were used to assess recognition modules' performance: Decidability (DEC) [4], Area Under Curve (AUC) and Equal Error Rate (EER). The evaluation of the stressed feature encoding techniques for the different working distances and traits is registered in Table 2. For a better interpretation of their performance, the Receiver Operating Characteristic (ROC) curves are also presented in Fig. 4. Results refer to a total of 69960 comparisons, performed in a 1:N fashion.

Table 2. Performance for each one of the exploited methods, traits and working distances. Metrics are Decidability (DEC), Area Under Curve (AUC) and Equal Error Rate (EER).

Trait →		Periocular						Face	Global
Method →		LBP	HOG	SIFT	ULBP	GIST	Fusion	PCA	Fusion
15m - 25m	DEC	0.802	0.699	0.404	1.090	0.918	1.162	1.171	1.407
	AUC	0.753	0.703	0.617	0.786	0.772	0.805	0.779	0.835
	EER	0.302	0.358	0.416	0.281	0.304	0.287	0.307	0.246
25m - 35m	DEC	0.677	0.641	0.341	0.972	0.808	1.033	1.173	1.267
	AUC	0.697	0.674	0.598	0.744	0.755	0.771	0.772	0.810
	EER	0.376	0.380	0.431	0.334	0.321	0.303	0.328	0.254
15m - 35m	DEC	0.529	0.520	0.310	0.830	0.747	0.891	0.676	1.025
	AUC	0.663	0.640	0.591	0.710	0.721	0.754	0.674	0.779
	EER	0.396	0.409	0.435	0.360	0.348	0.317	0.395	0.293

As we can see from Table 2, top recognition performance was attained at closer working distances (15m to 25m), with an AUC of 0.835. However, widening the working range to the whole driveway (15m to 35m), a considerable good performance is still achieved (AUC = 0.779). We must have in mind that results come from a fully automated system, operating on an adverse surveillance scenario. Furthermore, matches were not performed against a separate dataset of good registration images, but between different Pan-Tilt-Zoom (PTZ) images acquired during system operation.

As for the differences between the different exploited traits, the periocular region seems to be less affected by changes in distance, although further facial recognition

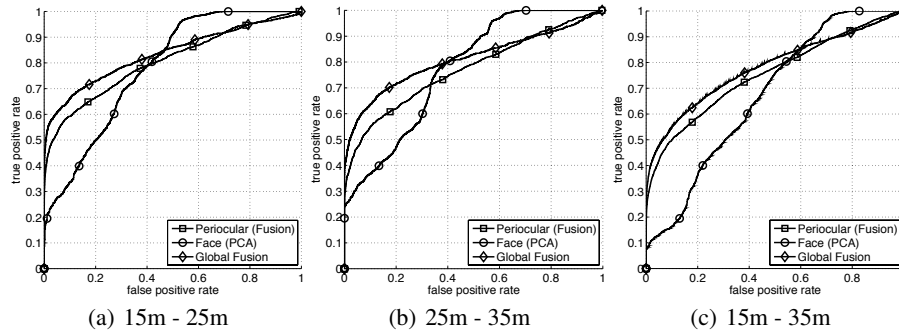


Figure 4. Receiver Operating Characteristic (ROC) curves for the periocular recognition, face recognition and global fusion, at different working distances.

techniques should be stressed. Also from the ROC curves at Fig. 4, we can see how the PCA applied to the face alone delivers lower true positive rate while introducing higher amounts of false positives, when compared to the fusion of methods operating on the periocular region. Nonetheless, fusing that information with the periocular methods scores produces a considerable improvement on the latter. Thus, if considering deploying a more restrictive system with higher security constraints, the face trait should not be used alone, but can be a powerful ally to further improve its final outcome.

4 Final Considerations

In this paper, we present the concept of a fully automated surveillance and biometric recognition system, able to complement human detection and tracking with biometric recognition over *in the wild* surveillance environments. Although further state-of-the-art techniques can be stressed for each module, we give evidence for the feasibility of such system, providing both tracking performance and biometric recognition results over a real surveillance scenario.

Although a functional system is presented, further work should be considered over three axes: 1) a larger dataset should be acquired, not only with a larger number of subjects going through the scene, but also with the system running over different environments (e.g. indoor lounge); 2) some modules are still to be developed, that would increase the recognition performance even further (e.g. head landmark detector); 3) additional state-of-the-art techniques should be tested for each module, and results cross-validated over the different scenarios. In particular, different face recognition techniques should be stressed, along with ear shape and iris biometrics and gait recognition.

References

1. S. Bharadwaj, H. Bhatt, M. Vatsa, and R. Singh. Periocular biometrics: When iris recognition fails. In *Fourth IEEE Int'l Conf. on Biometrics: Theory Applications and Systems (BTAS), 2010*, pages 1–6, Sep. 2010.

2. W. W. Bledsoe. The model method in facial recognition. Technical Report PRI 15, Panoramic Research, Inc., Palo Alto, California, 1964.
3. M. Castrillón, O. Denis, C. Guerra, and M. Hernández. Encara2: Real-time detection of multiple faces at different resolutions in video streams. *Journal of Visual Communication and Image Representation*, 18(2):130–140, 2007.
4. J. Daugman. High confidence visual recognition of persons by a test of statistical independence. *Pattern Analysis and Machine Intelligence*, 15(11):1148–1161, Nov. 1993.
5. I. Haritaoglu, D. Harwood, and L. Davis. W4: Real-time surveillance of people and their activities. *Pattern Analysis and Machine Intelligence*, 22(8):809–830, Aug. 2000.
6. A. Jain, S. Pankanti, S. Prabhakar, L. Hong, and A. Ross. Biometrics: A grand challenge. In *Proc. of the 17th Int'l Conf. on Pattern Recognition (ICPR)*, vol. 2, pages 935–942, 2004.
7. R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Trans. of the ASME – Journal of Basic Engineering*, (82 (Series D)):35–45, 1960.
8. B. Kamgar-Parsi, W. Lawson, and B. Kamgar-Parsi. Toward development of a face recognition system for watchlist surveillance. *Pattern Analysis and Machine Intelligence*, 33(10):1925–1933, Oct. 2011.
9. B. Keni and S. Rainer. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008.
10. L. Maddalena and A. Petrosino. A self-organizing approach to background subtraction for visual surveillance applications. *Image Processing, IEEE Trans.*, 17(7):1168–1177, July 2008.
11. J. Matey, O. Naroditsky, K. Hanna, R. Kolczynski, D. LoIacono, S. Mangru, M. Tinker, T. Zappia, and W. Zhao. Iris on the move: Acquisition of images for iris recognition in less constrained environments. In *Proc. of the IEEE*, vol. 94, pages 1936–1947, 2006.
12. U. Park, A. Ross, and A. Jain. Periocular biometrics in the visible spectrum: A feasibility study. In *IEEE 3rd Int'l Conf. on Biometrics: Theory, Applications, and Systems (BTAS)*, pages 1 – 6, Sep. 2009.
13. J. Shi and C. Tomasi. Good features to track. In *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, pages 593–600. IEEE, 1994.
14. C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, page 252, 1999.
15. M. Turk and A. Pentland. Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition, 1991. Proc. CVPR '91., IEEE Computer Society Conference on*, pages 586–591, 1991.
16. P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. of the 2001 IEEE Computer Society Conf. on Computer Vision and Pattern Recognition*, vol. 1, pages 511–518, 2001.
17. A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma. Toward a practical face recognition system: Robust alignment and illumination by sparse representation. *Pattern Analysis and Machine Intelligence*, 34(2):372–386, Feb. 2012.
18. W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, 2000.
19. Neves, J.C. and Moreno, J.C. and Barra, S. and Proença, H. A Calibration Algorithm for Multi-camera Visual Surveillance Systems Based on Single-View Metrology. In *Proceedings of the 7th Iberian Conference (IbPRIA)*, pages 552–559, 2015.
20. Neves, J.C. and Proença, H. Dynamic Camera Scheduling for Visual Surveillance in Crowded Scenes using Markov Random Fields. In *Proceedings of the 12th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS)*, 2015