# AVSS 2007: IEEE International Conference onAdvanced Video and Signal based Surveillance, London, UK, September 2007

Fihl, Preben

*Publication date:*
2009

*Document Version*
Publisher's PDF, also known as Version of record

Technical Report CVMT-09-01 ISSN 1601-3646, January 2009

# AVSS 2008: IEEE International Conference on Advanced Video and Signal based Surveillance, London, UK, September 2007

P. Fihl

Laboratory of Computer Vision and Media Technology

Aalborg University, Denmark

Email: pfa@cvmt.dk

***Summary:*** *This technical report will cover the participation in the IEEE International Conference on Advanced Video and Signal based Surveillance in September 2007. The report will give a concise description of the most relevant topics presented at the conference, focusing on the work related to the HERMES project and human motion and action recognition. Our contribution to the conference will also be described.*

# 1. General conference information

The fourth IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS) was held in London, UK, on 5-7 September 2007. AVSS is an annual conference with the location alternating between USA/Australia and Europe. The 2007 conference was hosted by Queen Mary, University of London.

The conference featured four keynote talks, three overview talks, nine oral sessions running as one track, and three poster sessions. The general scientific level of the oral sessions was high and the presented research was to a high degree relevant to the HERMES project and to human motion analysis and action recognition in general.

The conference offered two well organized social events which matched the scientific level and contributed to the impression of a successful conference. A welcome reception was given at the evening of the first conference day. The welcome reception was held on board the London Eye (a 135 meter high observation wheel located on the banks of the River Thames) were wines and soft drinks were served during two revolutions of the wheel. A banquet was held on the second conference day at the Octagon, an impressive facility at the Queen Mary, University of London. It was originally build as a library in 1887 but completely restored in 2006 as a flexible exhibition room.

# 2. Research from AVSS

The majority of the papers presented at AVSS relate to the research areas of the HERMES project and many papers also relates directly to human motion analysis and action recognition. This report will therefore present general ideas from a number of papers and only present details of the most interesting papers.

The conference covered many aspects of the area of automatic analysis of surveillance video and no general trend stands out when looking at the conference as a whole. However, the conference featured an industrial session and the number of papers that was authored or co-authored by employees of private companies indicate that automatic analysis of video is moving into the commercial world and from the presentations at AVSS the commercial video analysis systems seem to be quite successful.

Another noticeable thing about the presented research was the number of papers working with sensor networks in stead of a single sensor. The number and variety of sensors seem to be increasing, probably due to reduced sensor costs and increased computational power.

## 2.1 Invited talks

The first keynote talk was titled *Looking at People* and presented by Professor Dariu Gavrila from Daimler Chrysler Research & Technology and University of Amsterdam. The talk first presented the outcome of a large benchmark study on appearance-based person detection [1]. Next, the talk gave an overview of current research trends in model-based human motion analysis and discussed differences and similarities between model-based and appearance-based human motion analysis. The last part of the talk covered two

systems for “looking at people” – a real-time pedestrian detection system for crash avoidance in cars (with some remarkable experiments with real pedestrians and cars) and a system for aggression detection in public spaces. The talk was very interesting and entertaining and with a good scientific level.

The second keynote talk was given by Professor Tsuhan Chen from Carnegie Mellon University. The talk was titled *A journey from signal processing to surveillance* and discussed how signal processing has expanded from low-level processing to methods for high-level image understanding (e.g. SIFT features and Hidden Markov Models). The real interesting part of the talk however, was the application that was used to illustrate this issue, namely a system for object discovery. The system was capable of extracting the "object of interest" from a set of images in a completely unsupervised manner (under some general assumptions). The talk was equally interesting and entertaining to the first talk though not directly relevant to the HERMES project and to human motion analysis and action recognition.

The last two keynote talks were at the same high level as the first two but less relevant to my research. Andrew Blake from Microsoft Research Cambridge gave a talk on *Stereo Vision and Segmentation*. The talk presented both the development of a stereo web-camera and methods for real time background/foreground segmentation. The stereo web-camera has reached the market after the conference. The Hydra Stereo Webcam is sold by nVela (http://www.nvela.com/hydra.html) at prices from £250 and Microsoft Research Cambridge has released a SDK for the camera and their methods. The camera and SDK seem to be an interesting and rather cheap combination for working with stereo vision.

Dr. Tu from General Electric’s Global Research presented a talk on *Computer-aided Facial Reconstruction using Skulls.* The talk presented a joint General Electric/FBI project that generated computer renderings of faces based on the skull shape for crime investigations or missing person reports.

The most interesting of the three overview talks was by far the talk by John Garofolo from the U.S. National Institute of Standards and Technology (NIST) titled *Directions in Automatic Video Analysis Technology Evaluations at NIST*. The talk gave a very interesting discussion about the use of standard datasets for tests in computer vision. The talk presented how challenging standard datasets have played a significant role in speech recognition and parallels were drawn to the computer vision research. In speech recognition the first published results based on a new dataset often had recognition rates of 20-40% which would then improve over years were many researchers would use the same dataset. At some time a new dataset would be ready to give new challenges to the research community. In computer vision researchers rarely publish results with recognition rates below 75% and usually above 90%. The datasets are chosen or constructed so that high recognition rates are reachable. This results in many different datasets, each with its own characteristics suitable to get high recognition rates for a specific problem. The talk presented some efforts of NIST to produce challenging standard dataset and encouraged the computer vision research community to begin publishing results on such datasets and accepting low recognition rates on hard problems.

As a passing remark; an example of this problem was seen at ECCV 2008 and at the Workshop on Visual Surveillance 2008. The most dominant standard datasets within

action recognition are the KTH dataset and the Weizmann dataset. The rather simple Weizmann dataset is still widely used even though all new papers present recognition rates of 100% on this dataset.

## 2.2 Overview of research presented at AVSS

This section will give an overview of the research presented at AVSS and emphasize the research related to the HERMES project and the Ph.D. project.

The oral sessions on the first conference day covered sound surveillance, 3D face recognition and novel biometrics, and an industrial session.

The sound surveillance presentations showed some good results on detection of aggressive behavior, screams, and gunshots, and one paper did human identification based on the sound of people's gait. It seems very relevant to incorporate sound surveillance with video surveillance when sound recordings are feasible.

The face and biometrics session featured a 3D face recognition system, a method for compressing 3D face data, a method for pose normalization for face recognition, and a system for person identification based on 3D data of ears. The papers had a fine scientific level but none of them seemed to contain results with more than a slight impact.

The industrial session contained some nice presentations and the presented systems seemed quite impressive. However, only little detail was presented on the applied methods and most of the contributions stemmed from integration of several modules based on known methods or variations hereof.

The poster sessions on the first conference day covered sensor networks, object classification and recognition, activity monitoring and camera calibration, surveillance for transport systems, and target tracking. The most interesting papers from those sessions were: "Combination of Self-organization Map and Kernel Mutual Subspace Method for Video Surveillance" [2] and "Classifying and Tracking Multiple Persons for Proactive Surveillance of Mass Transport Systems"

[2] presents a method for classification of people and vehicles in far-field-of-view surveillance cameras. The method achieves good results but so does many other methods on people and vehicle classification. It is not obvious in what ways the method performs better. [3] presents a method for tracking and classification of multiple people. The tracking is done in two steps. First *blobs* are matched across frames based on color and area and next a particle filter is applied to track blobs through occlusions. The blob matching is quite similar to the work of Park and Aggarwal [4]. The classification of people is done based on edge templates. The novelty of the paper is limited and the tests are sparse but the methods are nicely combined and similar to the approach taken in my own research.

The oral sessions of the second conference day were: sensor fusion, video tracking, and the i-Lids challenge.

The first session presented three diverse sensor fusion papers. One paper fusing 2D and 3D images, i.e. color images and 3D data acquired by projecting a color pattern on to the

scene, one paper fusing sound and video and the last paper fusing RFID data with an acoustic sensor.

The video tracking session featured four papers. Two presented variations of known tracking methods. One used color-cues to enhance the re-sampling of particles in a particle filter and the other used an evaluation scheme of corner detections to aid tracking through occlusions. A third paper of the session used a complete 3D model of an office environment and two cameras to do tracking of multiple people, which became a rather simple task with the presented setup. The last paper presented an interesting notion of tracking guided by social information. The “social” information was based on similarity of different tracks, e.g. two people walking side by side would give two similar tracks and also indicate a social relation. This rather simple idea was however formulated in a quite general way that seemed to improve the tracking in terms of following people’s normal behaviors. The testing of the method was preliminary but showed promising results.

The i-Lids challenge aimed at detection of abandoned bags and parked cars in surveillance video. Seven papers were presented showing good results on the benchmark data sets of the challenge.

The poster sessions of the second conference day covered infrared imaging, military applications, gait and activity recognition, motion detection, and face localization and profiling. Our research was one of five posters of the gait and activity recognition session. However, the only interesting paper of that session was “View-invariant Human Feature Extraction for video-surveillance Applications” [5]. The other papers seemed to be preliminary work based on some very delimiting assumptions. From the other sessions the most interesting paper was “Fusion of Background Estimation Approaches for Motion Detection in Non-static Backgrounds” [6].

The view-invariant human feature extraction presented in [5] is based on a set of shape models where each shape model represents a human pose seen from 32 different view points. The paper use eight view points around the person and four levels of camera elevation for each of the eight view points. The shape models are trained in an offline process. In the online feature extraction the best view point is estimated from the direction of motion of the person and some ground plane estimation. The pose estimates are limited to people that are walking because of the training of the shape models. It would be intractable to use this approach for general pose estimation since the number of shape models would grow rapidly with even slightly advanced actions. This limitation is also in play in our gait analysis system and limits our work to gait.

[6] presented a fusion of a long term background model and a short term background model to improve motion detection. The method seemed efficient at handling dynamic backgrounds but no quantitative or comparative results were presented to support this claim.

The third and last conference day featured three oral sessions. They were multi-camera networks, video systems of retail applications, and vision-based human gesture and action recognition.

Papers on multi-camera networks and more generally multi-sensor networks were presented at many different sessions throughout the conference so the three papers of the multi-camera network session did not reflect the amount of research presented on that topic. The best paper from that session was “People Tracking Across Two Distant Self-calibrated Cameras” [7]. The paper presented a method for establishing correspondence between two distant cameras (with overlapping field-of-view) based on a combination of geometric properties of the scene and the appearance of people in the scene.

The session for video systems for retail applications covered four papers, three of these came from private companies. Interesting systems were presented, but with very limited levels of detail, which seem to be true for most industrial presentations. The focus was on the capabilities of the systems rather than the methods applied to achieve the presented results.

The last oral session on vision-based human gesture and action recognition featured some very interesting presentations. Three papers presented methods for sign language interpretation and hand gesture recognition. One paper presented a method for detecting and counting people in athletic videos [8] and the last paper presented a method for human pose estimation by fusion multiple cameras [9].

[8] presents good results on people detection and counting in challenging videos. A set of robust features are extracted from the silhouettes of people or groups of people. The features are the major axis angle, the eccentricity, and the silhouette area normalized with the estimated number of people in the scene.

[9] presents a framework for human pose estimation. One of the fundamental ideas of the systems is that people are being monitored by a camera network with small bandwidth. This implies local image processing at the cameras. The paper presents a method to model the human body by a set of ellipses mainly estimated from color-cues. These ellipses can then be transmitted over the network for fusion into a final 3D pose estimate. The method assumes that the projection of the 3D human model on to the 2D image plane is known. In the presented application this is achieved by knowing the camera position and the position of the person. The paper is conceptually very good but it is completely missing a results section (only one figure presents some pose estimates). This work will be interesting to follow in the future.

The poster sessions of the third conference day covered: face recognition, evaluation and description, statistical methods and learning, object-background segmentation, authentication and summarization, and target localization and tracking. Three papers stand out from these poster sessions as especially relevant to the HERMES project and the Ph.D. project: “3D Model-based People Detection and Tracking” [10], “Learning Gender from Human Gaits and Faces” [11], and “Representing and Recognizing Complex Events in Surveillance Applications” [12].

The 3D detection and tracking of people in [10] is done in a multi-camera setup. The cameras are calibrated and have overlapping fields of view. A simple cube is used as the 3D model for people and people are detected in each camera view by color-based foreground segmentation. The novelty of the paper seems very limited since standard methods are applied throughout the system. An underlying focus of the system seems to

be fast processing and usability of the system for non-expert users which may justify the system despite the limited novelty.

The gait analysis and face recognition in [11] is also based on known methods. The gait analysis is based on Gait Energy Images (a variant of Motion History Images and Motion Energy Images) while the face recognition is done based on three manually labeled features. The main contribution of the paper is the combination of the two approaches on the feature level. This is done using canonical correlation analysis. Good results are presented but the real contribution of the method is not clear since manually labeled features are included.

[12] is related to the conceptual level of the HERMES project. The paper uses the Semantic Web Rule Language (SWRL) to represent event in surveillance videos whereas HERMES uses Situation Graph Trees (SGT). The focus of the paper is to represent and process events in the format of SWRL, and the paper does not arrive at any results or conclusions that indicated whether or not the use of SWRL in stead of SGTs would benefit the HERMES project.

## 2.3 Own research at CRV

Our own research presented at AVSS was the paper “Classification of Gait Types Based on the Duty-factor”. It was part of the poster session titled “Gait and activity recognition”.

The paper deals with classification of human gait types based on the notion that different gait types are in fact different types of locomotion, i.e., running is not simply walking done faster. The paper presents the duty-factor, which is a descriptor based on this notion. The duty-factor is independent on the speed of the human, the cameras setup etc. and hence a robust descriptor for gait classification. The duty-factor is basically a matter of measuring the ground support of the feet with respect to the stride. This is estimated by comparing the incoming silhouettes to a database of silhouettes with known ground support. Silhouettes are extracted using the Codebook method and represented using Shape Contexts. The matching with database silhouettes is done using the Hungarian method. While manually estimated duty-factors show a clear classification the presented system contains misclassifications due to silhouette noise and ambiguities in the database silhouettes.

Not many comments were made regarding our work during the poster session.

# References

[1] S. Munder, D. M. Gavrila: *An Experimental Study on Pedestrian Classification*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.28, no.11, pp. 1863-1868, 2006.

[2] B. Zhang, J. Park, H. Ko: *Combination of self-organization map and kernel mutual subspace method for video surveillance.* IEEE Conference on Advanced Video and Signal Based Surveillance. 5-7 Sept. 2007

[3] S. Kong, C. Sanderson, B. C. Lovell: *Classifying and Tracking Multiple Persons for Proactive Surveillance of Mass Transport Systems*. IEEE Conference on Advanced Video and Signal Based Surveillance. 5-7 Sept. 2007

[4] Sangho Park and J.K. Aggarwal: *Segmentation and Tracking of Interacting Human Body Parts under Occlusion and Shadowing*, IEEE Workshop on Motion and Video Computing (WMVC 2002), Orlando, USA, 2002.

[5] G. Rogez, J. J. Guerrero, C. Orrite: *View-invariant Human Feature Extraction for video-surveillance Applications*. IEEE Conference on Advanced Video and Signal Based Surveillance. 5-7 Sept. 2007

[6] Eduardo Monari, Charlotte Pasqual: *Fusion of Background Estimation Approaches for Motion Detection in Non-static Backgrounds*. IEEE Conference on Advanced Video and Signal Based Surveillance. 5-7 Sept. 2007

[7] R. Pflugfelder, H. Bischof: *People Tracking Across Two Distant Self-calibrated Cameras*. IEEE Conference on Advanced Video and Signal Based Surveillance. 5-7 Sept. 2007

[8] C. Panagiotakis, E. Ramasso, G. Tziritas, M. Rombaut, D. Pellerin: *Automatic People Detection and Countning for Athletic Videos Classification*. IEEE Conference on Advanced Video and Signal Based Surveillance. 5-7 Sept. 2007

[9] C. Wu, H. Aghajan: *Model-based Human Posture Estimation for Gesture Analysis in an Opportunistic Fusion Smart Camera Network*. IEEE Conference on Advanced Video and Signal Based Surveillance. 5-7 Sept. 2007

[10] G. Garibotto: *3D Model-based People Detection and Tracking*. IEEE Conference on Advanced Video and Signal Based Surveillance. 5-7 Sept. 2007

[11] C. Shan, S. Gong, P. McOwan: *Learning Gender from Human Gaits and Faces*. IEEE Conference on Advanced Video and Signal Based Surveillance. 5-7 Sept. 2007

[12] L. Snidaro, M. Belluz, G. L. Foresti: *Representing and Recognizing Complex Events in Surveillance Applications*. IEEE Conference on Advanced Video and Signal Based Surveillance. 5-7 Sept. 2007

[13] P. Fihl, T. B. Moeslund: *Classification of Gait Types Based on the Duty-factor*. IEEE Conference on Advanced Video and Signal Based Surveillance. 5-7 Sept. 2007