

The AIT 2D Face Detection and Tracking System for CLEAR 2007

Andreas Stergiou, Ghassan Karame,
Aristodemos Pnevmatikakis and Lazaros Polymenakos

Athens Information Technology, Autonomic and Grid Computing,
P.O. Box 64, Markopoulou Ave., 19002 Peania, Greece
{aste, gkar, apne, lcp}@ait.edu.gr
<http://www.ait.edu.gr/research/RG1/overview.asp>

Abstract. This paper describes the AIT system for 2D face tracking and the results obtained in the CLEAR 2007 evaluations. The system is based on the complementary operation of a set of face detectors and a deterministic tracker based on color. To minimize false positives, the system is applied on the body regions provided by a stochastic body tracker, and utilizes a detection validation scheme based on color and texture modeling of the faces.

1 Introduction

Tracking and recognizing people is very important for applications such as surveillance, security and human-machine interfaces. In the visual modality, faces are the most commonly used cue for recognition. Finding the faces also helps resolve human bodies that are merged into one by the tracker. Hence face localization is of paramount importance in many applications.

Face localization can be done on a single camera frame by means of a detector, or across multiple frames using a tracker. Any of the two tasks can become very difficult in far-field unconstrained recording conditions. Low resolution faces suffering from pose, illumination and expression variations, as well as occlusions can only be detected sporadically. The misses need to be accounted for by a tracker, whose model needs frequent update to cope with the ever-changing face. Also, such face detectors suffer from false alarms that need to be constrained as much as possible.

Face detection can be very accurate [1,2], given large resolution, almost frontal pose and long processing time allowance. Unfortunately, none of these apply to the intended application, where resolution is low, pose can be arbitrary and processing has to be real-time. Cascades of simple classifiers [3] can detect small faces in arbitrary background and are fast. An ensemble of such cascades can be trained, each with different poses, together serving as a multi-view face detector.

Two approaches can be used in face tracking: stochastic and deterministic. Stochastic trackers are based on recursive Bayesian filtering, either in its exact form for Gaussian states and linear dynamics, the Kalman filter [4], or in its numerical approximation for non-linear dynamics, the particle filter [5]. Deterministic tracking

2 **Andreas Stergiou, Ghassan Karame,
Aristodemos Pneumatikakis and Lazaros Polymenakos**

on the other hand minimizes a cost function related to template matching between the object and a candidate region. In Mean-Shift [6] face tracking, or its continuously adaptive variant (CAM-Shift) [7] the template is a color histogram, that is very suitable for non-rigid objects. In both cases, the open question is the means and frequency of model update [8].

In [9], a system that utilizes a detector to find faces and a tracker to cope with misses operates in the whole image, increasing processing time and potentially false positives. In [10] foreground segmentation is also utilized. That system has been successfully tested in the Classification of Events, Activities and Relationships (CLEAR2006) evaluation [11,12].

In this paper, we enhance such systems in a number of ways. Three face detectors find frontal and left/right profile faces. Also, false alarms are minimized by applying the detectors on the body areas provided by a body tracker, by validating detections using a Gaussian Mixture Model (GMM) of face parameters and by merging frontal with profile detections.

The rest of the paper is organized as follows: In section 2, the face detection and tracking system is described, comprising the body tracker, the face detectors, the detection validation and merging scheme and the deterministic face tracker. In Section 3, the performance of the proposed system in the CLEAR 2007 evaluations [12] is analyzed. Finally, in Section 4 the conclusions are drawn.

2 Face Detection and Tracking System

The 2D face localization is constrained in the body areas provided by a body tracker. Three face detectors for frontal and left/right profile faces provide candidate face regions in the body areas. The face candidates are validated using the probability scores from a GMM. The surviving candidates are checked for possible merging, as both the profile detectors and the frontal one can detect different portions of the same face if the view is half-profile. The resulting face candidates are associated with faces existing in the previous frame and also with tracks that currently have no supporting evidence and are pending to either get an association, or be eliminated. Any faces of the previous frame that do not get associated with candidate faces at the current frame have a CAM-Shift tracker [7] initiated to attempt to track similarly colored regions in the current frame. If CAM-Shift also fails to track, then these past faces have their track in pending status for F_p frames. Finally, all active face tracks are checked for duplicates, i.e. high spatial similarity. The block diagram of the 2D face detection and tracking system is shown in Figure 1. In the following subsections, the various modules of the system are detailed.

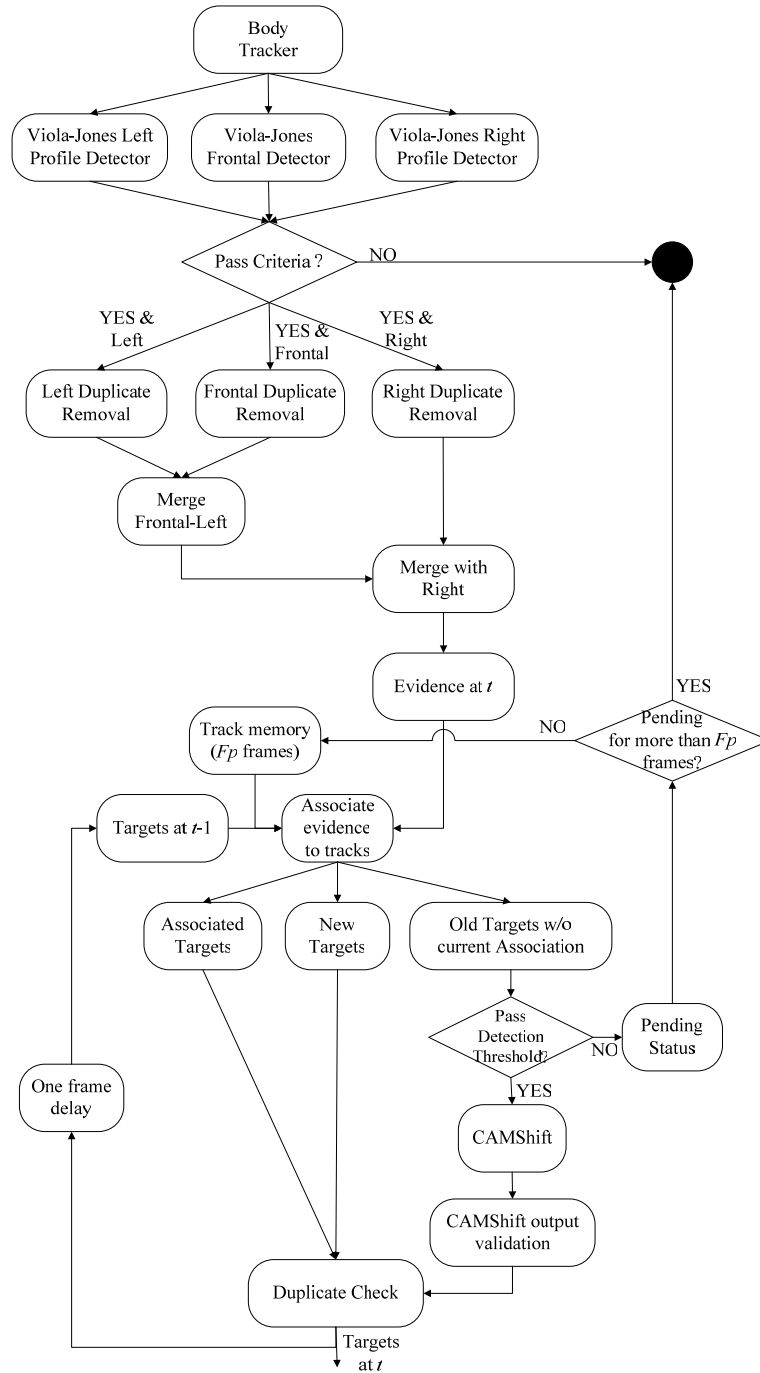


Fig. 1. Block diagram of the complete face localization system.

2.1 Body Tracking Module

The goal of the body tracker is to provide the frame regions occupied by human bodies. Any subsequent face detection and tracking is performed within these body regions. The tracker is based on a dynamic foreground segmentation algorithm [13,14] that utilizes adaptive background modeling with learning rates spatiotemporally controlled by the states of a Kalman filter [4]. It comprises three modules in a feedback configuration: adaptive background modeling based on Stauffer's algorithm [15] provides the pixels that are considered foreground to the evidence formation module. The latter combines the pixels into body evidence blobs, used for the measurement update state of the Kalman filter module. The states of the Kalman filter are used to obtain an indication of the mobility of each target, as a combination of translation motion and its size change. Also the position and size of the targets are contained in the states of the Kalman filter. This information is fed back to the adaptive background modeling module to adapt the learning rate in the vicinity of the targets: frame regions that at a specific time have a slow-moving target have smaller learning rates.

The proposed spatiotemporal adaptation of the learning rate of the adaptive background modeling module solves the problem of Stauffer's algorithm when foreground objects stop moving. Without it, targets that have stopped moving are learnt into the background. With the proposed feedback configuration this process is halted long enough for the intended application, i.e. tracking people in a meeting. For the details of the body tracking algorithm and its application both for in-doors and out-doors tracking, see [13,14].

2.2 Face Detection and Validation Module

The goal of the face detection and validation module is to find possible faces in a given frame without the use of any temporal information, and to validate them in order to reduce the false detections. Many face detectors can be found in the literature [1-3]. They are split into those that attempt to find face features, and from them find the faces themselves, and those that search directly for faces. The former are usually slower, and require adequate resolution for accurate feature detection. As the faces are small compared to the frame size, the natural choice for a detector is the boosted cascade of simple features [3]. Its implementation in OpenCV [16] is chosen, as this is publicly available. We train a left profile and a frontal detector. Right-profiles are detected by flipping the frame and using the left-profile detector. For each detector we use 6,000 positive samples (images with marked faces), 20,000 negative samples (images with no human or animal face present), an aspect ratio of 3/4, minimum feature size 0, 99.9% hit rate and 50% false alarm per cascade stage, horizontal and 45-degrees tilted features, non-symmetric faces and gentle AdaBoost learning [16]. The detector thus trained is applied on the grayscale portions of the frames that are designated as human bodies by the body tracker. Only the upper part of the bodies that has height equal to the body width is examined by the detector to speed up the process.

Unfortunately the face detector suffers from false detections. These are constrained

by the fact that the detector is applied on a limited portion of the frames, but nevertheless the detections need to be validated to further reduce false alarms. The face validation module is based on a multivariate Gaussian Mixture Model comprising representative statistics of skin color and brightness of the detected region. The models of Jones and Rehg [17] are used to build the likelihood ratio of human skin versus non-skin. Although detected faces are for the most part expected to contain areas with high skin color likelihood, we need also to account for regions surrounding the eyes and eyebrows that do not exhibit skin-like colors. Furthermore, human faces exhibit a lot of brightness variation due to self-shadowing, and can thus not be uniformly too bright or too dark. Such conditions are implicitly enforced by the GMM, so that false positives on furniture, walls and, to some extent, hand regions can be removed. The likelihood threshold for face validation used by the detectors is determined during the training stage of the GMM so as to correctly validate 99% of the training faces, and is usually somewhat different for frontal and profile faces.

Frequently, the various face detectors yield multiple detections of the same face target, thus resulting in duplicate tracked objects. Such a behavior could be especially detrimental when the face in question is somewhat rotated, either in- or out-of-plane. We remedy this by combining such duplicate detections. That is, matching frontal and left profile detections are first merged into single face targets. The resulting objects are subsequently combined with matching right profile detections. Such a strategy aims at improving the target extension accuracy by preventing direct combination of left and right profile detections. Detections are marked as matching for merging when they meet two conditions: they have to be located within a distance twice the maximum width of the detections in question and the resulting merged detection should have both width and height smaller than a threshold of 80 pixels.

2.3 Face Tracking Module

After possible merging, the detected faces are assigned to the tracked targets. This association is done using an optimal greedy algorithm, the Hungarian (or Munkres) algorithm [18], which minimizes the overall Mahalanobis distance between the detections and the targets. Once the detections are assigned to some targets, the targets' records are updated accordingly. On the other hand, in case a detection can not be assigned to any target, a new target is initialized. Finally, if a target is not associated to any detection at a certain frame, for instance due to occlusion, rotation or tilt, then the target's record is examined. If it has more detections associated with it than a threshold, then deterministic tracking based on histogram matching and using the CAM-Shift algorithm [7] is initiated. Otherwise, the target in question enters pending status. The threshold is a decreasing linear function of the age of the target; older targets need smaller detection rates to be kept active.

The CAM-Shift tracker uses a trained color histogram in RGB color space as a representation of the face to be tracked. The training of the color histogram is carried out in the area of the last detection. The histogram is not updated throughout tracking; only the detections that are validated by the GMM are trusted for the task. The pixels that are used for the histogram are the top 75 percentile that match skin color according to the Jones and Rehg skin color model. Utilizing the trained histogram and

a probability threshold of 0.05, a binary map of the pixels in a search region around the face location in the previous frame is built. Higher probability thresholds lead more constraint tracked regions. Morphological closing removes any small regions on the binary map. Then the centroid and width of the updated face location is calculated using moments [7]. The height is computed using the aspect ratio of the detected face that initiated the tracker. The CAM-Shift tracked regions are also validated using the same GMM as the regions returned by the detectors; only the threshold is relaxed to account for the looser face framing evident in most CAM-Shift tracked regions.

When the CAM-Shift tracker does not yield any face region, then the track enters pending status. This mechanism accounts for temporary occlusions of the face under tracking; e.g. when the person in question is not facing the camera for some frames. In this case, the target's history is kept in memory for Fp frames, after which, it is erased. If, on the other hand, a detection occurs in the vicinity of the pending track before it is erased, then the track becomes active again. Hence the pending mechanism allows for track continuity while it prevents false alarm faces from being reported.

Finally, all active face tracks are checked for duplicates, in order to prevent duplicate tracking of the same face. The check for target duplicates utilizes both target location and area.

3 CLEAR 2007 Evaluation Results

The proposed two-dimensional face localization system is tested on the CLEAR 2007 evaluations, on multi-camera indoor video sequences coming from the CHIL. Typical results are shown in Figure 2.

The quantitative evaluation of the proposed system follows the CLEAR2007 evaluation protocol [11]. According to it, the tracking system outputs (hypotheses) are mapped to annotated ground truths based on centroid distance and using the Hungarian algorithm [18]. The metrics for face tracking are five [11]. The Multiple Object Tracking Precision (MOTP) is the position error for all correctly tracked persons over all frames. It is a measure of how well the system performs when it actually finds the face. There are three kinds of errors for the tracker, false positives, misses and track identity mismatches. They are reported independently and also jointly in an accuracy metric, the Multiple Object Tracking Accuracy (MOTA). The MOTA is the residual of the sum of these three error rates from unity. The quantitative performance of the system is summarized in Table 1.

It is evident from Table 1 that the easiest recording site is AIT. This is due to the small size of the room that allows easier detection. Two sites suffer particularly from misses: ITC and UPC. The reasons are different. No faces from ITC have been used in training the detectors, since in the 2006 recordings only the presenter was tracked. Most faces in the ITC recordings are severely tilted, so it is difficult to initiate and maintain face tracks. The main reason for misses in the UPC recordings is the color of the faces: shadows and darker skin colors cause some faces to fail the detection validation. The severe interlacing of the moving people is a secondary cause.

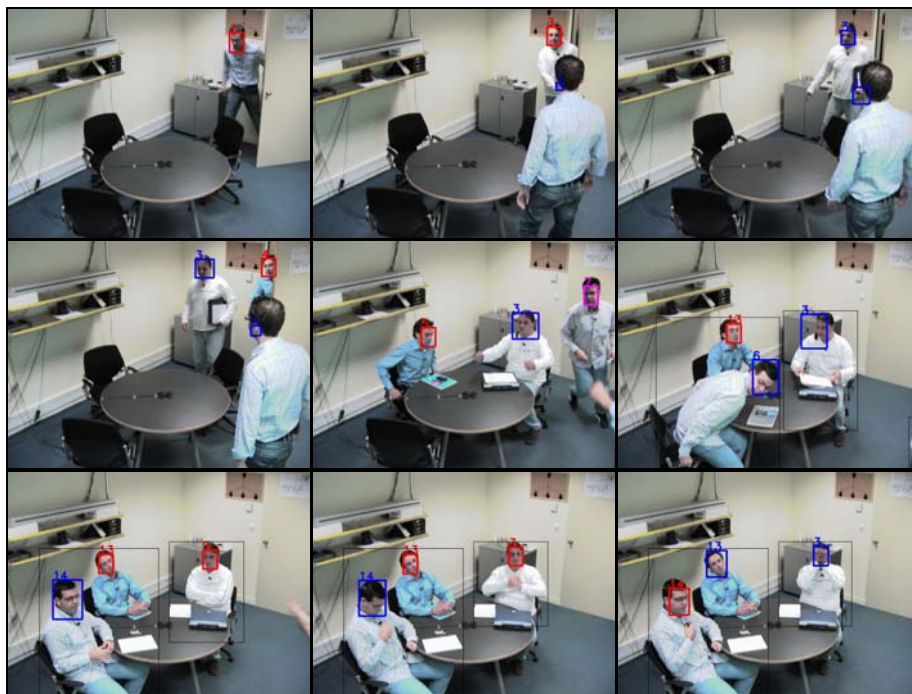


Fig. 2. Typical performance of the face localization system. Detections of the three cascades of simple classifiers are marked in red, while faces being tracked by the CAM-Shift tracker are marked in blue. Notice that the latter are occluded or tilted faces.

Table 1. Face detection performance of the system on all 40 seminars and 4 camera views, averaged per site.

Site	MOTP	Misses (%)	False positives (%)	Mismatches (%)	MOTA (%)
AIT	0.665	34.0	13.1	1.64	51.3
IBM	0.672	36.7	35.2	2.06	26.1
ITC	0.723	65.5	13.2	1.62	19.7
UKA	0.610	33.5	31.2	3.45	31.9
UPC	0.656	50.0	12.1	2.32	35.6
Overall	0.656	42.5	22.1	2.29	33.1

Two sites suffer from many false positives: IBM and UKA. For the IBM recordings, the poor performance of the GMM validation scheme is due to the skin-like table-top. On the other hand, for most of the UKA recordings that scheme was turned off entirely. This is due to the ‘blue’ dominant color in those recordings that renders the general human skin color model unusable. The hit rate versus the false positive rate for each of the 40 seminars is shown in Figure 3. The points are grouped per each of the five recording sites. The false positive rate is constrained mostly

below 20% for three sites: AIT, UPC and ITC. While for AIT the hit rate is mostly high, for UPC it varies a lot per seminar and for ITC it is mostly too low. On the other hand the hit rate of IBM and UKA does not vary a lot per seminar, being high for UKA and moderate for IBM. This is due to the very small face sizes in the IBM recordings. On the other hand, for these sites the false positive rate varies a lot per seminar.

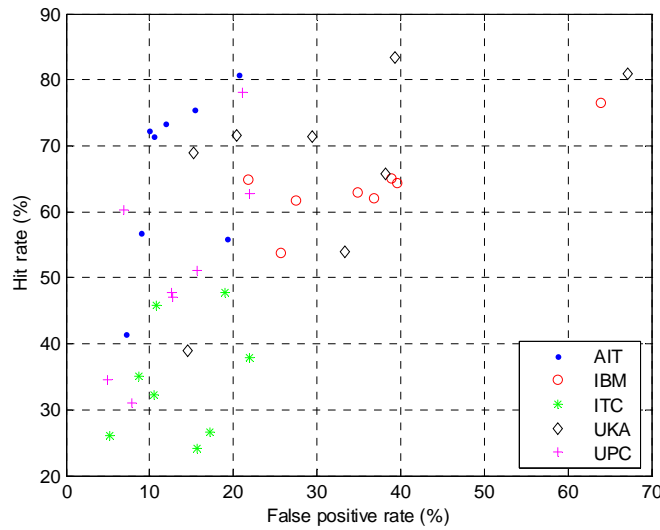
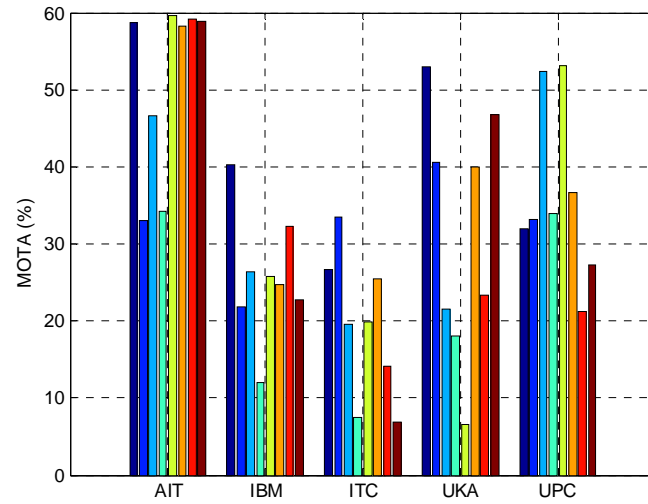


Fig. 3. Hit versus false alarm rate for each of the recording five sites and eight seminars per site.

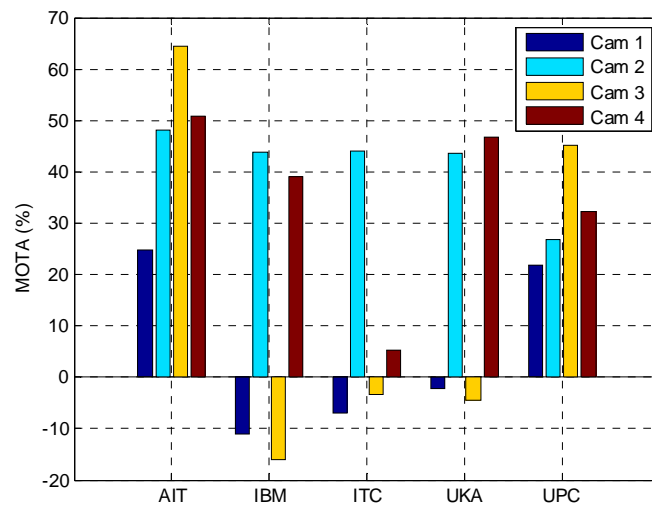
It is instructive to explore the variation of the MOTA of the system across the five different sites and four different recording segments per site. This is done in Figure 4.a. The variation of performance per camera in each site is also important. This is shown in Figure 4.b. Obviously, cameras 1 and 3 are problematic in the IBM, ITC and UKA recordings. This does not affect performance greatly, since these are the camera views with the smallest number of ground truth faces.

4 Conclusions

In the proposed face localization system, a CAM-Shift tracker complements three detectors based on cascades of simple features. The CLEAR evaluations have shown that the system should be improved by reducing misses and false positives. Both can be achieved employing better detectors, based on FloatBoost [19] or Schneiderman's detector [20,21]. Also, better modeling of faces versus non-face patches can help eliminate false positives. Finally, the performance of the overall system should be assessed when the CAM-Shift tracker is replaced by a Kalman or particle filter one.



(a)



(b)

Fig. 4. Break-down of MOTA performance in each recording site per seminar (a) and camera (b).

Acknowledgements

This work is sponsored by the European Union under the integrated project CHIL, contract number 506909. The authors wish to thank the people involved in data collection, annotation and overall organization of the CLEAR 2007 evaluations for providing such a rich test-bed for the presented algorithm.

References

- [1] S.-Z. Li and J. Lu: Face Detection, Alignment and Recognition, in G. Medioni and S. Kang (eds.), *Emerging Topics in Computer Vision*, (2004).
- [2] R.-L. Hsu, M. Abdel-Mottaleb and A. K. Jain: Face Detection in Color Images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24 (2002), 696-706.
- [3] P. Viola and M. Jones: Rapid Object Detection using a Boosted Cascade of Simple Features, *IEEE Conf. on Computer Vision and Pattern Recognition*, (Dec 2001), 511.
- [4] R. E. Kalman: A New Approach to Linear Filtering and Prediction Problems, *Transactions of the ASME – Journal of Basic Engineering*, 82 (Series D), (1960) 35-45.
- [5] S. Arulampalam, S. Maskell, N. Gordon and T. Clapp: A Tutorial on Particle Filters for On-line Non-linear Non-Gaussian Bayesian Tracking, *IEEE Transactions on Signal Processing*, 20 (2002), 174-188.
- [6] G. Jaffré and A. Crouzil: Non-rigid object localization from color model using mean shift, in *International Conference on Image Processing (ICIP 2003)*, (Sept. 2003).
- [7] G. Bradski: Computer Vision Face Tracking for Use in a Perceptual User Interface, *Intel Technology Journal*, 2 (1998).
- [8] S. Zhou, R. Chellappa and B. Moghaddam: Visual tracking and recognition using appearance-adaptive models in particle filters, *IEEE Transactions on Image Processing*, 13 (2004), 1491-1506.
- [9] T. Yang, S.-Z. Li, Q. Pan, J. Li and C. Zhao: Reliable and Fast Tracking of Faces under Varying Pose, in *7th International Conference on Automatic Face and Gesture Recognition (FGR 2006)*, (Apr. 2006), 421-426.
- [10] K. Bernardin, T. Gehrig and R. Stiefelhagen: Multi- and Single View Multiperson tracking for Smart Room Environments, in R. Stiefelhagen and J. Garofolo (eds.) *CLEAR 2006, Lecture Notes in Computer Science*, 4122 (2007), 81-92.
- [11] R. Stiefelhagen, K. Bernardin, R. Bowers, J. Garofolo, D. Mostefa and P. Soundararajan: The CLEAR 2006 Evaluation, in R. Stiefelhagen and J. Garofolo (eds.) *CLEAR 2006, Lecture Notes in Computer Science*, 4122 (2007), 1-44.
- [12] www.clear-evaluation.org
- [13] A. Pnevmatikakis and L. Polymenakos: Kalman Tracking with Target Feedback on Adaptive Background Learning, in S. Renals, S. Bengio and J. Fiscus (eds.), *MLMI 2006, Lecture Notes in Computer Science*, (2007).
- [14] A. Pnevmatikakis and L. Polymenakos: Robust Estimation of Background for Fixed Cameras, in *International Conference on Computing (CIC2006)*, (2006).
- [15] C. Stauffer and W.E.L. Grimson: Learning patterns of activity using real-time tracking, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22 (2000), 747-757.
- [16] G. Bradski, A. Kaehler and V. Pisarevsky: Learning-Based Computer Vision with Intel's Open Source Computer Vision Library, *Intel Technology Journal*, 9 (2005).

- [17] M. Jones and J. Rehg: Statistical color models with application to skin detection, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, (1999), 274–280.
- [18] S. Blackman: Multiple-Target Tracking with Radar Applications, Artech House, Dedham, MA (1986), chapter 14.
- [19] S.-Z. Li and Z.Q. Zhang: FloatBoost Learning and Statistical Face Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26 (2004), 1112-1123.
- [20] H. Schneiderman: Feature-Centric Evaluation for Efficient Cascaded Object Detection, in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, (June 2004).
- [21] M. Nechyba and H. Schneiderman: PittPatt Face Detection and Tracking for the CLEAR 2006 Evaluation, in R. Stiefelhagen and J. Garofolo (eds.) *CLEAR 2006, Lecture Notes in Computer Science*, 4122 (2007), 161-170.