

The AIT Face Tracker for VACE Multisite Meeting Recordings

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 VACE multisite meeting recordings section of the CLEAR 2007 evaluations. The system is based on the complementary operation of a set of face detectors and a tracker. To minimize false positives, the system utilizes a detection validation scheme based on color.

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 on the other hand minimizes a cost function related to template matching between the

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

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 this paper, we use different face detectors to find frontal and profile faces. Limitations of resolution and the presence of large pose, expression and illumination variations make face detection very difficult. To be able to obtain faces the operating point of the detectors is set to allow false positives. Hence a face validation scheme is introduced, to eliminate as many of them as possible. Also, a merging scheme merges frontal with profile detections that belong to the same face. Finally, a tracker accounts for the misses of the detectors. A Kalman tracker that utilizes the color information of the face is initiated after every detection that is not associated with a new detection in the next frame. Similar approaches of complementing the face detector with a tracker can be found in [9,10].

The rest of the paper is organized as follows: In section 2, the face detection and tracking system is described, comprising the face detectors, the detection validation and merging schemes and the stochastic face tracker. In Section 3, the performance of the proposed system in the VACE multisite meeting recordings section of the CLEAR 2007 evaluations [11,12] is analyzed. Finally, in Section 4 the conclusions are drawn.

2 Face Detection and Tracking System

The localization of faces can in principle be constrained in the body areas provided by a body tracker. Although this approach is followed in [13], it is not used for the VACE multisite meeting recordings section of the CLEAR 2007 evaluations as there is not enough motion in most of these videos for the AIT body tracker [14,15] to initiate the tracks. Three face detectors for frontal and left/right profile faces provide candidate face regions by searching in the entire frame. The face candidates are validated using a skinness map, i.e. the probability of the colors in the regions to correspond to human skin. 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 Kalman tracker initiated to attempt to track similarly colored regions in the current frame. If the tracker also fails, 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 face detection and tracking system for the VACE multisite meeting recordings section of the CLEAR 2007 evaluations is shown in Figure 1. In the following subsections, the various modules of the system are detailed.

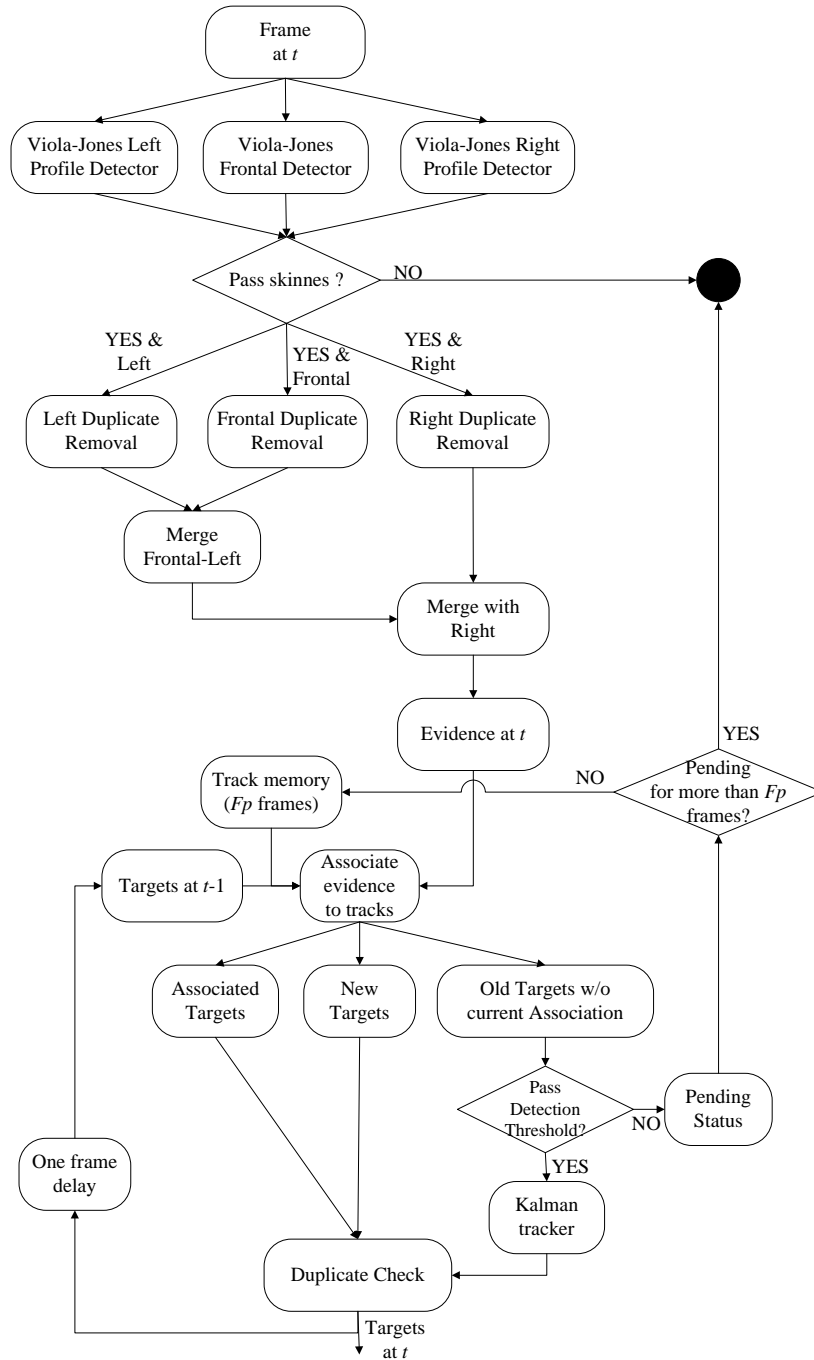


Fig. 1. Block diagram of the complete face localization system.

2.1 Face Detection Module

The goal of the face detection module is to find possible faces in a given frame without the use of any temporal information. 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 the training of 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 marked faces that are used as positive samples are obtained from the annotated faces of the CLEAR 2006 test sequences. The detectors thus trained are applied on the grayscale frames.

2.2 Face Validation Module

Unfortunately the face detector suffers from false detections. Hence the detections need to be validated to reduce false alarms. The face validation module is based on a skin likelihood map that enumerates the similarity of the colors in the detected face regions to human skin. The models of Jones and Rehg [17] are used to build the likelihood ratio of any color RGB triplet belonging to human skin versus non-skin. The skin likelihood map corresponding to the whole frame is thresholded, leading to skin-colored patches. To validate a detected face region, then within its bounding box there needs to be a significant percentage of pixels that are active in the thresholded skin likelihood map.

Note that in [13] we have used a different detection validation scheme, based on Gaussian Mixture Modeling of color and texture properties of the faces marked in the training sequences of the CHIL face tracking task [12]. This is not possible for the VACE multisite meeting recordings section of the CLEAR 2007 evaluations as no part of the provided sequences, although termed ‘training’ has the faces annotated to extract the models from.

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 Kalman tracking 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 Kalman tracker uses a constant velocity model and evidence collected from the frame by thresholding a likelihood map that enumerates the similarity of the RGB color triplets of the pixels to the colors found in the face being tracked. To do so, a color histogram in RGB color space is trained as a representation of the face to be tracked. The training of the color histogram is carried out using the pixels of the regions detected by the face detectors. The histogram is not updated throughout tracking; only the detections that are validated are trusted for the task. The histogram update uses memory, placing more confidence to the most recent detections.

In [13], we used a deterministic CAM-Shift tracker instead of the Kalman one. The main reason for trying to change the tracker in this work is because we have not been very satisfied with the framing of the faces by the CAM-Shift tracker. Also, in [13] the histogram of a face is trained in using only the latest detection, not a memory of all past detections associated with the particular track.

When the 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 VACE CLEAR 2007 Evaluation Results

The proposed two-dimensional face localization system is tested on the VACE multisite meeting recordings section of the CLEAR 2007 evaluations. Typical results are shown in Figure 2.

The quantitative evaluation of the proposed system follows the VACE evaluation protocol [19]. According to it, the tracking system outputs (hypotheses) are mapped to annotated ground truths. The primary metrics for face tracking are two: 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 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 detection versions of these two metrics, the Multiple Object Detection Precision (MODP) and Multiple Object Detection Accuracy (MODA) do not consider track identity mismatches. The quantitative performance of the system is summarized in Table 1.

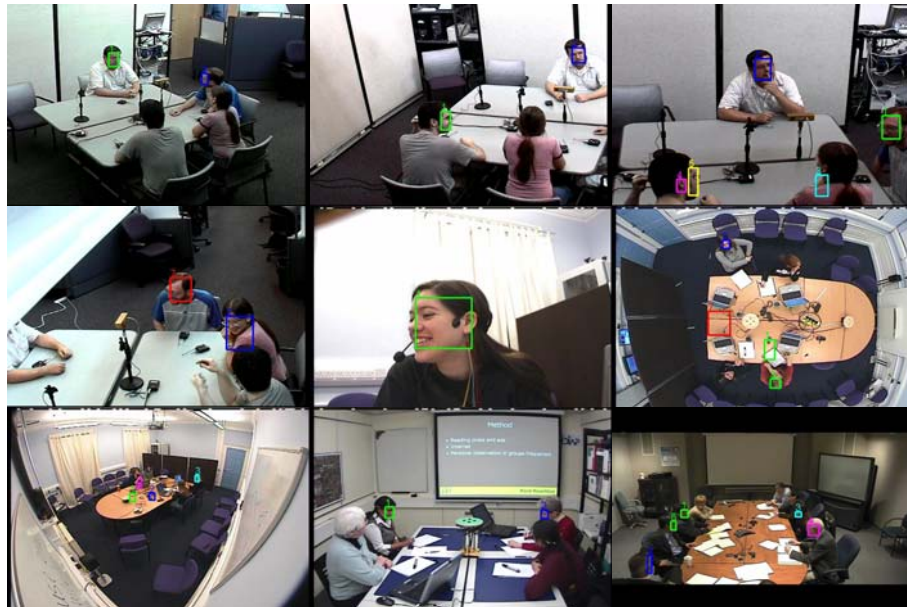


Fig. 2. Typical performance of the face localization system. Note the huge scale differences between the different videos. Also note the false positives introduced by the portions of the background that have skin-like color (like the light brown tabletops). Finally note the problems introduced by ceiling cameras.

**The AIT Face Tracker
for VACE Multisite Meeting Recordings 7**

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

Site	Camera	MOTA	MOTP	MODA	MODP
CMU	1	0.754117	0.389168	0.760345	0.39108
	2	0.610005	0.471459	0.613843	0.481971
	3	0.10886	0.470489	0.113903	0.481669
	4	0.066071	0.272116	0.094017	0.267079
	1	0.854083	0.356301	0.862888	0.362882
	2	0.164589	0.574879	0.170599	0.594912
	3	-0.33799	0.493185	-0.33201	0.510621
	4	0.685068	0.547175	0.687984	0.556083
EDI	1	0.825202	0.391323	0.891429	0.391323
	2	0.917533	0.356488	0.956667	0.356488
	3	0.96356	0.539589	0.985795	0.539589
	4	0.826019	0.469796	0.839599	0.469796
	6	-0.46557	0.518283	-0.46358	0.520069
	1	0.915931	0.555009	0.944251	0.555009
	2	0.857769	0.442454	0.912913	0.442454
	3	1	0.533721	1	0.533721
IDI	5	-1.09589	0.444647	-1.09246	0.438951
	6	-3.89232	0.367822	-3.88942	0.360969
	1	0.880132	0.374985	0.88716	0.374985
	2	1	0.608063	1	0.608063
	4	0.8125	0.598335	0.8125	0.598335
	6	0.341924	0.605769	0.341924	0.605839
NIST	7	-0.18336	0.605199	-0.18287	0.616589
	1	-51.9672	0.547779	-51.9672	0.576739
	2	-0.41419	0.480783	-0.41419	0.481809
	6	0.529101	0.478348	0.542627	0.49683
	7	0.463855	0.486226	0.463855	0.486226
	2	0.57988	0.562361	0.589633	0.565895
	3	0.37452	0.575775	0.375184	0.56464
	6	0.93782	0.649301	0.93782	0.647581
TNO	7	0.998567	0.548702	1	0.548702
	1	1	Not defined	1	Not defined
	5	-1.62702	0.367315	-1.62559	0.361337
	5	-0.12322	0.297492	-0.12322	0.295899
VT	6	0.352629	0.459642	0.353383	0.443571
	2	0.273996	0.423164	0.279825	0.427221
	3	0.242871	0.553586	0.243547	0.554587
	4	-0.24323	0.256464	-0.22691	0.256464
	5	0.613096	0.436159	0.621802	0.430904
	6	0.003308	0.354638	0.021053	0.36074
	1	0.058727	0.339733	0.062937	0.33558
	2	-0.34652	0.336181	-0.33235	0.348369
	3	0.636678	0.446199	0.636678	0.451405
	5	0.355096	0.43923	0.364042	0.444365
7	0.404306	0.430923	0.411417	0.433509	
Median values		0.39122	0.47049	0.39599	0.48167

The MOTA per site are presented in Figure 3. It is evident from that and Table 1 that the EDI and IDI are the easiest sites, CMU, NIST and VT are about as hard as all

the sites pooled together, and TNO is the most difficult. The reason for EDI and IDI being the easier sites is because those recordings have many near-field cameras. On the other hand, the difficult TNO recordings have skin-like colors in the background, rendering the rejection of false positives useless. This ranking is based on the median values, but due to the variation of the results across the different cameras and the limited number of recordings analyzed per site, there is large uncertainty about the estimation of the median values, rendering the ordering of the difficulty statistically risky.

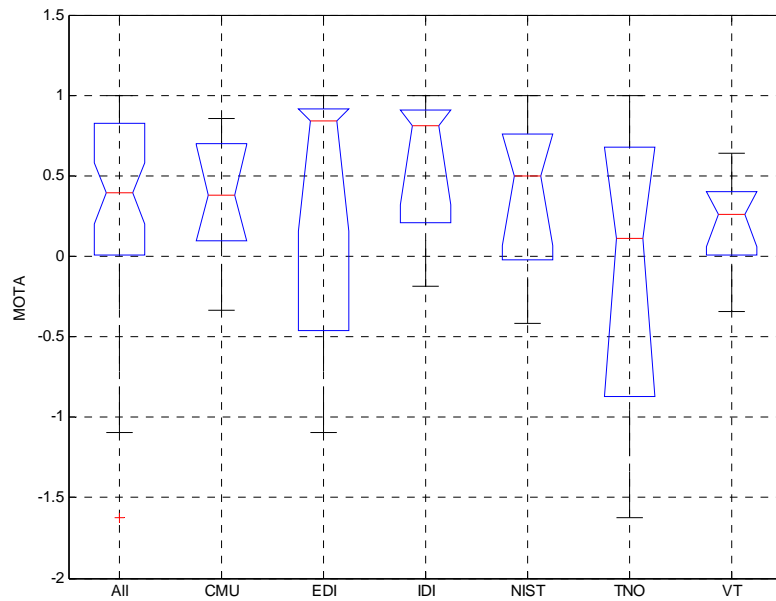


Fig. 3. Boxplot of the MOTA of all the sites pooled together, and site-specific. EDI has the highest median MOTA, but also the largest variation and uncertainty in the estimation of the median.

4 Conclusions

In the proposed face localization system, a Kalman 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 Kalman tracker is replaced by a CAM-Shift or particle filter one.

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. Stergiou, G. Karame, A. Pnevmatikakis and L. Polymenakos: The AIT 2D face detection and tracking system for CLEAR2007, in R. Stiefelhagen and J. Garofolo (eds.) *CLEAR 2007*, submitted.
- [14] 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).
- [15] A. Pnevmatikakis and L. Polymenakos: Robust Estimation of Background for Fixed Cameras, in *International Conference on Computing (CIC2006)*, (2006).
- [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).

10 **Andreas Stergiou, Ghassan Karame,
Aristodemos Pnevmatikakis and Lazaros Polymenakos**

- [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] R. Kasturi, et. al: Performance evaluation protocol for face, person and vehicle detection & tracking in video analysis and content extraction (VACE-II), University of South Florida (Jan 2006).
- [20] 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.
- [21] H. Schneiderman: Feature-Centric Evaluation for Efficient Cascaded Object Detection, in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, (June 2004).
- [22] 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.