3D Tracking of Multiple People Using Their 2D Face Locations

Nikos Katsarakis , Aristodemos Pnevmatikakis and Michael Nechyba
Athens Information Technology, Autonomic and Grid Computing Group
0.8km Markopoulou Ave., PO Box 68, 19002 Peania, Greece
{nkat, apne}@ait.edu.gr
http://www.ait.edu.gr/research/RG1/overview.asp
Pittsburgh Pattern Recognition, 40 24th Street, Suite 240, Pittsburgh,
PA 15222, USA
michael@pittpatt.com
http://www. pittpatt.com

Abstract. In this paper, we address tracking of multiple people in complex 3D scenes, using multiple calibrated and synchronized far-field recordings. Our approach utilizes the faces detected in every camera view. Faces of the same person seen from the different cameras are associated by first finding all possible associations and then choosing the best option by means of a 3D stochastic tracker. The performance of the proposed system is evaluated by using the outputs of two grossly different 2D face detectors as input to our 3D algorithm. The multi-camera videos employed come from the CLEAR evaluation campaign. Even though the two 2D face detectors have very different performance, the 3D tracking performance of our system remains practically unchanged.

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.

Tracking in smart rooms can be very complicated due to the complex background and the crowded foreground. In smart room scenarios the cameras are placed a bit above body height, aiming to have their optical axes almost perpendicular to the faces and they are not panoramic, in order to have reasonable face sizes. This results to viewing conditions where a small group of four people can fill up a significant part of the frame. In such cases the targets no longer contain a

Please use the following format when citing this chapter:

Katsarakis, N., Pnevmatikakis, A., Nechyba, M., 2007, in IFIP International Federation for Information Processing, Volume 247, Artificial Intelligence and Innovations 2007: From Theory to Applications, eds. Boukis, C., Pnevmatikakis, I., Polymenakos, I., (Boston: Springer), pp. 365-373.

single person. As the bodies in the image plane touch each other, a formed target contains them all.

Resolving the problem using motion or color information is not generic enough; the bodies can have similar velocity vectors, and they can be dressed in similar colors, especially in military applications. What can help to resolve the problem is the existence of multiple cameras; information from all of them can be used to build a 3D understanding of the smart room, enabling 3D person tracking.

In this paper we address person tracking in 3D by utilizing the 2D face locations on multiple calibrated [1] and synchronized cameras. Effectively, we track the centroid of the head of every person in multi-view and multi-person recordings. Possible associations of the different views of a face are constructed by projecting a grid of 3D points onto the different image planes and collecting face evidence. A stochastic tracker then selects the best association.

The recordings of the CLEAR Evaluation (Classification of Events, Activities and Relationships) [2,3] are utilized to test the proposed system. These recordings comprise five cameras each (four at the room corners and one panoramic at the ceiling). They depict multiple people in cluttered backgrounds, recorded in five different sites. The situation recorded is of the business meeting with presentation type. Breaks where people move around the room a lot are also recorded. These recordings are quite challenging for tracking.

The paper is organized as follows: In section 2 the 2D face trackers employed are outlined. In section 3 the 3D tracking system is detailed. In section 4 the results are discussed, to be followed by the conclusions in Section 5.

2 Face tracking in 2D

Two different 2D face tracking algorithms are utilized to produce the location of the faces in every camera view. The one is from Athens Information Technology (AIT) and the other from Pittsburgh Pattern Recognition (PittPatt). Both algorithms participated in the 2D face tracking task of the CLEAR 2007 evaluation campaign, giving quite different results. They are summarized in the next two subsections. Their performance is also presented in a third subsection

2.1 AIT face tracker

The AIT face tracker is detailed in [4]. In summary, it operates as follows: 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 Gaussian Mixture Model. 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 [5] 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 a predefined number of frames. Finally, all active face tracks are checked for duplicates, i.e. high spatial similarity. Typical results of the face tracker are shown in Fig. 1.

Fig. 1. Typical performance of the AIT face tracker. 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

2.2 PittPatt face tracker

The PiiPatt face tracker is detailed in [6]. In summary, it proceeds in three stages: (1) frame-based face detection; (2) motion-based tracking; and (3) track filtering. At the heart of this system lies PittPatt's robust face detector, available for single-image testing through a web demo [7]. Conceptually, this version of the detection algorithm builds on the approach developed by Schneiderman [8]; however, a number of changes have been implemented recently that dramatically boost speed performance over previous versions. These changes include code-level as well as algorithmic optimizations, and have led to better than real-time processing of video on contemporary PC platforms. In motion-based tracking, single-frame observations are combined into face tracks -- each of which is ultimately associated with a unique subject ID -- by exploiting the spatiotemporal continuity of video through a second-order motion model. Finally, in track filtering, results are finalized by merging IDs

for partial face tracks that meet certain spatial consistency criteria, and by eliminating face tracks likely to be false alarms. False alarm tracks are most often characterized by low classifier confidence throughout and/or exhibit very little movement throughout the lifetime of the track. It is these tracks that are eliminated in this last stage of processing.

2.3 Performance of the 2D face trackers

The quantitative evaluation of the both face trackers follows the CLEAR 2007 evaluation protocol [2,3]. According to it, the tracking system outputs (hypotheses) are mapped to annotated ground truths based on centroid distance and using the Hungarian algorithm [9]. The ground truths contain both the face bounding boxes and the number of fiducial points visible in the face. There such fiducial points are marked: the left and right eyes and the nose bridge. A marked face is considered of interest if at least two of the fiducial points are visible; faces with just one fiducial point are considered 'do not care' regions. The metrics for face tracking are five [2]. 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. There are 20 recordings, 4 from each recording site, employed in the evaluation. Each of them is 5 minutes long and comprises 4 corner cameras that are used for face tracking, and a fifth that is only optionally used in 3D tracking. The quantitative performance of the two systems is summarized in Tables 1 and 2.

Table 1. Per site and overall performance of the AIT face tracking system in the CLEAR 2007 multi-site and multi-camera recordings

Site	MOTP	MOTA	False positives	Misses	ID switches
AIT	0.67	46.52	13.1	33.99	6.39
IBM	0.66	16.46	42.66	34.93	5.96
ITC-IRST	0.68	-1.80	59.97	38.69	3.14
UKA	0.60	26.64	39.59	30.07	3.69
UPC	0.64	25.15	30.66	36.68	7.51
Overall	0.64	23.46	36.25	34.67	5.61

The two systems have a large difference in MOTA performance, mainly due to their difference in misses and false alarms. The difference in misses and part of the difference in false alarms is due to the superior performance of the Schneiderman face detector [8] employed in the PittPatt system, over the AdaBoost face detector [10] employed in the AIT system. A significant part though of the false positives is due to the color-based tracking using CAM-Shift [5] employed in the AIT system. This tracking allows the survival of targets that once have been faces, but then they have out-of-plane rotated, offering to the camera just some skin patch in the back of

the neck/head or the cheeks. For examples, see the tracked skin patches in the third, fourth and ninth frames of Fig. 1.

Table 2. Per site and overall performance of the PittPatt face tracking system in the CLEAR 2007 multi-site and multi-camera recordings

Site	MOTP	MOTA	False positives	Misses	ID switches
AIT	0.68	77.39	11.21	10.06	1.34
IBM	0.70	58.04	5.19	36.20	0.57
ITC-IRST	0.65	65.92	12.16	21.14	0.78
UKA	0.70	65.38	18.49	15.23	0.91
UPC	0.67	77.94	7.20	13.38	1.48
Overall	0.68	68.81	10.31	19.85	1.03

While such patches contribute to false positives to the face tracking tasks, they are quite useful to the 3D head tracking task for which the 2D face tracks are to be used in the next section. In order to demonstrate this difference in the two face tracking systems, they are both evaluated including the faces with just one fiducial point marked on them. The overall results for both systems are shown in Table 3. Obviously the performance of the AIT system degrades more gracefully as these faces are included.

Table 3. Overall performance of both face tracking system in the CLEAR 2007 multi-site and multi-camera recordings when face patches with just one fiducial point are included in the evaluation

Site	MOTP	MOTA	False positives	Misses	ID switches
AIT	0.63	27.49	24.76	41.95	5.80
PittPatt	0.67	62.83	6.97	29.04	1.16

3 Head tracking in 3D

Our approach for 3D tracking utilizes the 2D face localization system presented in the previous section, applied on multiple calibrated [1] and synchronized cameras. To solve the problem of associating the views of the face of the same person from the different cameras, a 3D space to 2D image planes approach is utilized. The space is spanned by a 3D grid. Each point of the grid is projected onto the different image planes. Faces whose centers are close to the projected points are associated to the particular 3D point. 3D points that have more than one face associated to them are used to form possible associations of views of the face of the same person from the different cameras. If in each camera view c there are n_c faces then the k-th association (of the total K ones) that span the 3D space is of the form $a^{(k)} = \{i_1^{(k)}, \ldots, i_C^{(k)}\}$, where C is the number of available cameras and $i_c^{(k)} \in \{0,1,\ldots n_c\}$. A value $i_c^{(k)} = 0$ corresponds to no face from the c-th camera in the k-th association, while any other value corresponds to the membership of a face from

those in the c-th camera in the k-th association. Obviously $\forall c \in \{1,\ldots,C\}$, $i_c^{(k_1)} > 0$ and $k_1 \neq k_2$, it is $i_c^{(k_1)} \neq i_c^{(k_2)}$, i.e. the same face in a camera view cannot be a member of different valid associations This condition renders some of the associations mutually exclusive. After eliminating duplicate associations, the remaining ones are grouped into possible sets of mutually exclusive associations and sorted according to a weight that depends on the distance of each association from the face center and on the number of other associations that contradict it.

All the M mutually exclusive sets of possible associations $a^{(k)}$ are validated using a Kalman filter in the 3D space. For each new frame, all possible solutions are compared to the state established on the previous frame, penalizing solutions which fail to detect previously existing targets, or in which there are detections of new targets in the scene. While this strategy reduces the misses and false positives, it does not prevent new targets from appearing, as in the case of new people entering the room, all solution pairs will include that new target and thus will be equally penalized.

The recordings employed also offer a fifth camera, a panoramic one. Although this camera can not be used for face tracking, it is quite useful for 3D head tracking. The AIT body tracker [11] is employed to obtain body bounding boxes from the panoramic camera. Any head being tracked by the 3D system should be included in this bounding box. If it is not, then the 3D head track is actually the product of miss-association of actual 2D face tracks, or correct association of false positives 2D tracks. Therefore the panoramic camera is used to verify the associations and thus improve the accuracy of the 3D system.

4 Performance evaluation

3D person tracking is defined in the CLEAR evaluations as [2] tracking of the projection of the head centroid on the floor. Qualitatively, the performance of the 3D tracking system is shown in Fig. 2. For the quantitative performance analysis, the same metrics used in 2D face tracking are utilized here as well. The results are sown in Tables 4 to 7, where the 3D tracker operates on the 2D face tracking results of the AIT system without (Table 4) or with (Table 5) the use of the panoramic camera validation and on the 2D face tracking results of the PittPatt system without (Table 6) or with (Table 7) the use of the panoramic camera validation.

Table 4. Per site and overall performance of the proposed 3D tracking system operating on the 2D faces provided by the AIT face tracker, without the use of the panoramic camera validation

Site	MOTP (mm)	MOTA (%)	False positives	Misses	ID switches
AIT	84.10	59.91	3.06	34.90	2.12
IBM	94.40	61.19	6.81	30.72	1.28
ITC-IRST	102.1	59.22	7.15	31.53	2.10
UKA	96.50	46.97	5.15	45.77	2.11
UPC	91.20	65.19	7.16	25.61	2.04
Overall	94.08	58.37	6.06	33.66	1.90

Table 5. Per site and overall performance of the proposed 3D tracking system operating on the 2D faces provided by the AIT face tracker, with the use of the panoramic camera validation

Site	MOTP (mm)	MOTA (%)	False positives	Misses	ID switches
AIT	83.70	59.23	2.42	36.18	2.16
IBM	93.40	62.06	5.27	31.39	1.28
ITC-IRST	86.10	62.82	2.56	32.79	1.83
UKA	96.20	47.90	2.62	47.36	2.12
UPC	89.40	67.74	4.03	26.22	2.02
Overall	90.52	59.91	3.52	34.71	1.86

Table 6. Per site and overall performance of the proposed 3D tracking system operating on the 2D faces provided by the PittPatt face tracker, without the use of the panoramic camera validation

Site	MOTP (mm)	MOTA (%)	False positives	Misses	ID switches
AIT	79.10	66.99	2.59	28.43	2.00
IBM	92.60	36.12	3.45	58.47	1.97
ITC-IRST	85.40	62.98	2.85	32.18	1.99
UKA	100.2	47.31	4.40	46.25	2.05
UPC	76.00	77.81	2.02	18.67	1.50
Overall	87.39	57.21	3.11	37.80	1.89

Table 7. Per site and overall performance of the proposed 3D tracking system operating on the 2D faces provided by the PittPatt face tracker, with the use of the panoramic camera validation

Site	MOTP (mm)	MOTA (%)	False positives	Misses	ID switches
AIT	77.50	66.21	1.84	29.95	2.00
IBM	92.30	37.50	1.50	59.02	1.97
ITC-IRST	83.70	60.51	1.34	36.36	1.78
UKA	97.30	46.45	2.29	49.25	2.01
UPC	75.40	78.29	0.94	19.26	1.50
Overall	86.01	56.91	1.57	39.68	1.84

It is evident from the results that the averaged performance of the 3D tracker is similar, no matter which 2D face tracks are used, with the use of the AIT 2D tracker together with the panoramic camera validation scheme yielding somewhat better MOTA. Since the AIT and PittPatt 2D face trackers yield grossly different results, this seems a counterintuitive result. The reason for this is the way the two 2D face trackers function: The AIT system allows for skin-colored head patches to be tracked, while the PittPatt system does not. As a result, the AIT 2D face tracker has reduced performance for faces, but enhanced for heads, yielding more frequent head detections from more than one camera to be synthesized into 3D tracks by the 3D system. The stricter face tracks of the PittPatt 2D system result to less frequent head detections from more than one camera, hence to more difficult 3D associations. Examining the results per recording site, two of them have similar performance no matter the 2D tracker employed, two are somewhat better with the PittPatt 2D tracker, and another one is far better with the AIT 2D tracker. Not surprisingly, in the

IBM recordings in which the use of the PittPatt 2D tracker does not give good 3D tracking results, the PittPatt 2D tracker has increased misses, possibly due to the very small face sizes. In terms of precision, the use of the PittPatt 2D face tracker improves (reduces) MOTP.

Fig. 2. Operation of the individual 2D face trackers on the four corner cameras and association of the 2D evidence into 3D tracks. The detected faces are marked by bounding boxes. The IDs of the tracks are of the form AIT_XXX shown at the projection of the tracked head centroids on the floor. The tracks are also projected to a panoramic camera (not used by the system) for better visualization

A second observation has to do with the effectiveness of the panoramic camera validation. When the AIT 2D face tracks are utilized by the 3D tracker, this validation scheme yields some improvement of the overall MOTA. In particular, for the recordings suffering from high false positive rates, the validation scheme considerably reduces the false positives, at the expense of some increase of the misses. Overall, the MOTA is increased. On the other hand, regarding the utilization of the PittPatt 2D face tracks, the false positives are lower, leading to no room for drastic improvement with the application of the validation scheme. Since the misses are again increased, the overall MOTA decreases.

5 Conclusions

In this paper we have proposed a 3D head tracker for cluttered scenes and have evaluated its performance according to the CLEAR evaluation protocol. The tracker utilizes 2D face tracks obtained from synchronized and calibrated cameras. Two such 2D face tracking systems have been employed in the evaluation, demonstrating the robustness of the proposed 3D tracker. The 2D systems from AIT and PittPatt employed in the evaluation have grossly different face tracking performance, but their output combined in 3D head tracks by our system results to similar 3D performance.

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 algorithms.

References

1. Z. Zhang: A Flexible New Technique for Camera Calibration, Technical Report MSR-TR-98-71, Microsoft Research, (Aug. 2002).
2. 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.
3. www.clear-evaluation.org
4. Stergiou, G. Karame, A. Pnevmatikakis and L. Polymenakos: The AIT 2D face detection and tracking system for CLEAR 2007, in R. Stiefelhagen, R. Bowers and J. Garofolo (eds.) *CLEAR 2007, Lecture Notes in Computer Science*, accepted.
5. G. Bradski: Computer Vision Face Tracking for Use in a Perceptual User Interface, *Intel Technology Journal*, 2, (1998).
6. M. Nechyba, L. Brandy and H. Schneiderman: PittPatt Face Detection and Tracking for the CLEAR 2007 Evaluation, in R. Stiefelhagen, R. Bowers and J. Garofolo (eds.) *CLEAR 2007, Lecture Notes in Computer Science*, accepted.
7. http://demo.pittpaftt.com
8. H. Schneiderman: Feature-Centric Evaluation for Efficient Cascaded Object Detection, in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, (June 2004).
9. S. Blackman: Multiple-Target Tracking with Radar Applications, Artech House, Dedham, MA (1986), chapter 14.
10. G. Bradski, A. Kaehler and V. Pisarevsky: Learning-Based Computer Vision with Intel's Open Source Computer Vision Library, *Intel Technology Journal*, 9, (2005).
11. Pnevmatikakis and L. Polymenakos: Robust Estimation of Background for Fixed Cameras, in *International Conference on Computing (CIC2006)*, (2006).