PERSON TRACKING IN SMART ROOMS USING DYNAMIC PROGRAMMING AND ADAPTIVE SUBSPACE LEARNING

Zhenqiu Zhang[†,1]*, Gerasimos Potamianos,*[2] *Stephen M. Chu*[2]*, Jilin Tu*[1]*, and Thomas S. Huang*[1]

[1] Beckman Institute, University of Illinois, Urbana, IL 61801, USA
[2] IBM T.J. Watson Research Center, Yorktown Heights, NY 10598, USA
Emails: zzhang6@uiuc.edu, {gpotam,schu}@us.ibm.com, {jilintu,huang}@ifp.uiuc.edu

ABSTRACT

We present a robust vision system for single person tracking inside a smart room using multiple synchronized, calibrated, stationary cameras. The system consists of two main components, namely initialization and tracking, assisted by an additional component that detects tracking drift. The main novelty lies in the adaptive tracking mechanism that is based on subspace learning of the tracked person appearance in selected two-dimensional camera views. The subspace is learned on the fly, during tracking, but in contrast to the traditional literature approach, an additional "forgetting" mechanism is introduced, as a means to reduce drifting. The proposed algorithm replaces mean-shift tracking, previously employed in our work. By combining the proposed technique with a robust initialization component that is based on face detection and spatio-temporal dynamic programming, the resulting vision system significantly outperforms previously reported systems for the task of tracking the seminar presenter in data collected as part of the CHIL project.

1. INTRODUCTION

Visual tracking of humans in complex scenes is a very interesting and challenging problem, a major difficulty being the non-stationary visual target characteristics, due to view, pose, or illumination changes. Various target representations have been used in the literature for this purpose, such as parameterized shapes [1], color distributions [2], image templates [3] and the eigen-space approach [4], to name a few. Tracking with fixed representations however is not reliable over long durations, and a successful tracker needs to allow appropriate model adaptation. Not surprisingly, a number of tracking methods have been developed to allow such adaptation, for example the online EM-algorithm based technique of [5], the online feature selection mechanism of [6], and the parametric statistical on-line appearance modeling technique in [7]. An interesting non-parametric approach appears in Lim et al. [8], where the appearance subspace is learned online by an efficient sequential algorithm for principal component analysis (PCA), updated employing the incoming data vectors. The proposed algorithm has been demonstrated to track human motion under large lighting and pose variations at almost real-time. However, the approach utilizes all input observations up to the present time, with no mechanism to discard distant past ones, that may have little in common to the current lighting, pose, and view conditions.

In this paper, we propose a modification to this scheme, that allows a forgetting mechanism to discard what we view as distant past input to the appearance modeling approach. We test this technique in a task that is of interest for human-computer interaction in smart rooms, namely that of tracking the 3D location of a person giving a lecture inside a room equipped with multiple multi-modal sensors. This task constitutes one of the focal technologies developed under the CHIL project [9]. The particular domain provides realistic labeled data, but also allows for additional information to detect and avoid tracking drift, thus easing the evaluation of our proposed approach. Such information is available through time-synchronous views from multiple calibrated cameras.

In this paper, we integrate the proposed adaptive subspace learning within a complete 3D single person tracking system in the CHIL lecture scenario. The system, in addition to the proposed tracking algorithm that operates on 2D camera views independently, has an initialization component which is based on face detection in the four camera views, triangulation, and dynamic programming. Furthermore, a drifting mechanism controls the switching between initialization and tracking. The proposed system constitutes a complete redesign of our previously reported approach [10], by replacing the motion-based initialization component with the dynamic programming framework introduced in [11], and the previously used mean-shift tracking technique with the proposed adaptive subspace methodology. Evaluation results on the CHIL database demonstrate dramatic performance improvements.

The rest of the paper is organized as follows: Section 2 presents an overview of the whole system, with the initialization and tracking components described in Sections 3 and 4, respectively. Experiments are presented in Section 5, and a brief summary in Section 6 concludes the paper.

2. SYSTEM OVERVIEW

As discussed above, in this paper we concentrate on 3D single person tracking in the CHIL seminar scenario. In this particular domain, multiple synchronized calibrated cameras are set up in a smart room, among them four corner room cameras with widely overlapping fields of view. A schematic of such room is depicted in Fig. 1. In our work, the inputs of the four corner cameras are used to obtain over time the 3D head position of a person presenting a seminar in front of an audience inside this room.

The overview diagram of the proposed system is given in Fig. 2(a). As already mentioned, it consists of an initialization and a tracking component, with tracking drift detection controlling the switch between these two modes. Briefly, for initialization, *multi-view face detectors* are first applied to four camera views in the smart room. Subsequently, spatio-temporal information of the face detection results over 10 consecutive quad-frames is integrated within a *dynamic programming* (DP) framework, to provide robust initialization. Details are described in Section 3 (see also Fig. 2(b)). If the optimal DP trajectory is accepted as a true object, the tracking component kicks in, operating independently in two camera views, selected from the four views, based on the DP result. Details of the tracking stage are presented in Section 4. As long as the DP trajec-

[†] This work was performed while Zhenqiu Zhang was on a summer internship with the Human Language Technology Department at the IBM T.J. Watson Research Center, and was supported by the European Commission under the integrated project CHIL, "Computers in the Human Interaction Loop", contract number 506909.

1-4244-0367-7/06/$20.00 ©2006 IEEE

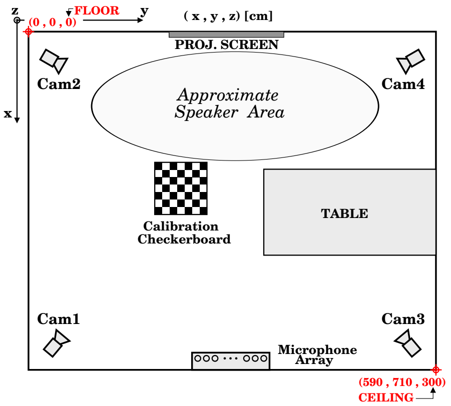

Fig. 1. Schematic diagram of the CHIL smart room at the Universität Karlsruhe, Germany.

tory is not acceptable, the initialization process is repeated with a shift of five frames.

An important aspect of the system is the re-initialization decision, or equivalently, the drift detection. This is based on a combination of local face detection and calibration-based triangulation to test the consistency of independent tracking in the two (selected based on the DP results) camera views. In more detail, if the inter-ray distance of the two 2D-to-3D mapping rays is larger than a predetermined threshold, this indicates that the two tracked results are inconsistent, hence immediately prompting re-initialization. Furthermore, at each frame, multi-view face detectors (see next Section) are also applied around the tracking result to determine whether there indeed exists a face object in the local region (in our system, this is an 80×80 pixel region). If faces could not be detected in the local region for several frames (30 in our case) in any of the two camera views, a re-initialization decision is prompted.

3. SPATIO-TEMPORAL FACE DETECTION FOR INITIALIZATION

For automatic initialization, the system is primarily based on multi-view face detectors, trained using the FloatBoost approach [12], on four camera images. In particular, two face detectors are trained: One for the frontal view and the other for the left side view (the right side face detector is obtained by mirroring the latter), both trained as a cascade, multi-layer structure of weak classifiers. For example, for our experiments (see Section 5), 15 layers and 576 features are used for the frontal face detector, and 30 layers with 4330 features for the left (see also [10]).

However, face detection by itself produces rather poor results in the challenging CHIL domain considered. This is illustrated in Fig. 3: The resolution of the presenter's face in each camera view is small, around 30×30 or less within the 640×480-pixel camera views, with significant pose and illumination change in the video sequence. Robust multi-view face detection in this scenario is really hard, with high rates of missing face detections and false alarms observed.

To solve this problem, we have proposed a novel algorithm in [11] to integrate spatial and temporal information available within the multi-camera video sequence setting. This replaces our previously employed motion-based framework [10]. In summary (see also Fig. 2(b)), the trained multi-view face detectors are first applied on all four camera views. Based on the spatial consistency of the detection result from different camera views, 3D hypotheses of the presenter's head location are generated using the calibration information. Then dynamic programming (DP) on the results over ten

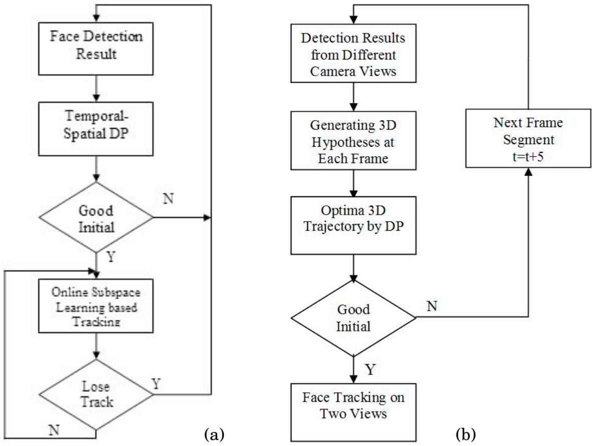
Face Detection Result
Temporal-Spatial DP
Good Initial
N
Y
Online Subspace Learning based Tracking
Lose Track
Y
N
(a)
Detection Results from Different Camera Views
Generating 3D Hypotheses at Each Frame
Optima 3D Trajectory by DP
Good Initial
N
Y
Face Tracking on Two Views
Next Frame Segment t=t+5
(b)

Fig. 2. Block diagram of the proposed multi-camera tracking system. (a) Overview; (b) Initialization.

consecutive frames is used to search for the optimal trajectory of the presenter's face in the 3D space, based on local similarity measure and a transition cost. If the optimal trajectory is accepted compared to a threshold, the result is fed into the tracking component described in the next section; otherwise the process is iterated with a five frame shift until an acceptable trajectory is determined. Some details are given next:

Generating 3D Hypotheses: Assuming n_i face detections per camera view, there could be

$$1/2 \times \sum_{i,j:i \neq j} n_i \times n_j \tag{1}$$

3D candidate head locations, obtained via triangulation. Based on the resulting inter-ray distances of the 2D-to-3D maps, one can easily reject few large inter-ray distance hypotheses. An additional height threshold is also set to distinguish the seminar presenter from the typically sitting audiences members. The remaining hypotheses are used in the DP framework.

Generating Optimal Dynamic Programming Trajectory: The DP framework contains three main components:

Local Similarity Measure: This is used to evaluate the hypothesis at the current instant on basis of the available four camera views. The color histograms of rectangles (approximately double the face height) in different views are used for this task, with the Bhattacharyya coefficient employed over 30-bin histograms of the H component of the color HSV space. The assumption is that if the candidate hypothesis is a true target, then the corresponding rectangles in different camera views should cover the same person, and color histogram similarity should be high.

Transition Cost: This penalizes non-smooth trajectories. In our framework, the transition is defined as the 3D spatial distance between two hypotheses, with its cost specified using a Gaussian diffusion with a pre-set diagonal covariance matrix [11]. A constant new trajectory generation cost is also defined.

Hypothesis Search: The searching scheme employs the standard dynamic programming framework, as described in [11]. A few things to note: A total of six hypotheses are kept "alive" at each time instant, as a pruning mechanism; a maximum acceptable score (constant) is set, thus providing a mechanism to reject the final hypothesis

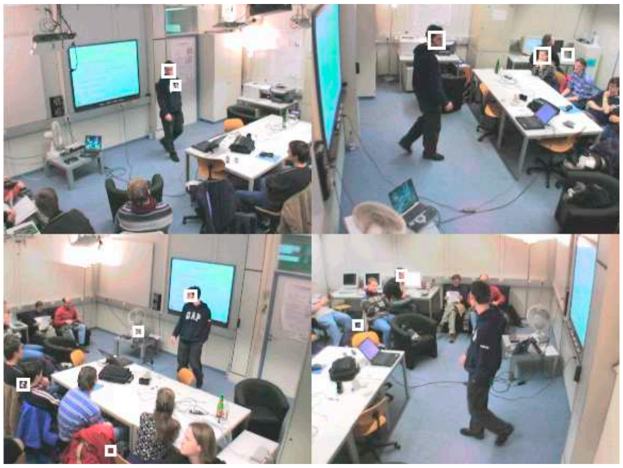

Fig. 3. Detection result on four synchronized camera views.

(and hence trigger a new search – with a five quad-frame shift); and finally, that the returned optimal trajectory defines the two camera views on which tracking is to commence, based on the views that generated the last-instant optimal trajectory hypothesis.

4. ADAPTIVE SUBSPACE TRACKING WITH A FORGETTING MECHANISM

In [8], when a new observation is obtained, the PCA subspace is updated to take into consideration the variance contributed by the new observation. However, the method does not provide an updating algorithm for eliminating past observations during tracking. This poses a problem when tracking video over long periods of time, as the noise introduced during tracking would eventually bias the PCA subspace away from the characteristic appearance of the desired tracked object. In [13], an L_∞ norm subspace is fitted to the past K frames incrementally by Gramm-Schmitt orthogonalization. Though the subspace with L_∞ norm has the advantage of incorporating observation novelties into the subspace representation in a timely manner, as shown by many successful experiments [13], it runs the risk of tracking drift, as consistent noise and outliers presented in the observation may easily bias the subspace away from the object appearance space. Considering that PCA offers the freedom for the user to perform dimensionality reduction to ignore the tracking noise and achieve outlier rejection by reconstruction error [4], we adopt the incremental PCA subspace learning approach, with Hall's mechanism [14] to incrementally update the PCA subspace given new observations. Furthermore, this algorithm allows us to adjust the subspace by eliminating distant past observations in the subspace. This introduces a forgetting mechanism that is absent in Lim's approach [8].

Suppose the subspace at time t is represented by $\mathcal{C}_t = (\bar{x}_t, U_t, \Lambda_t, N_t)$, where $\bar{x}_t$ is the mean vector, U_t is the eigenmatrix, Λ_t is the eigenvalue matrix, and N_t records the current number of observations modeled by the subspace. The incremental subspace learning algorithm is listed as Alg. 4.1 when a new observation y_t is presented, and the incremental subspace forgetting algorithm is listed as Alg. 4.2 when a past observation z needs to be eliminated.

In our particular implementation, we use the most recent 50 frame observations to construct the PCA subspace. Hence, after tracking initialization, the forgetting mechanism does not commence until after 50 frames are observed. For this initial duration, the algorithm remains the same as [8]. The learned subspace has dimensionality 15, down from the 20×20-pixel data "template".

Algorithm 4.1: INCRE_LEARN($\bar{x}_{t-1}, U_{t-1}, \Lambda_{t-1}, N_{t-1}, y_t$)

```
$N_t = N_{t-1} + 1$
$\bar{x}_t = \frac{\bar{x}_{t-1} N_{t-1} + y_t}{N_t}$
$d = \bar{x}_{t-1} - y_t$
$g = U'_{t-1} d$
$h = d - Ug$
if $||h|| = 0$
$A = \Lambda_{t-1} \frac{N_{t-1}}{N_t} + gg' \frac{N_{t-1}}{N_t^2}$
  else
  $\begin{cases} v = h/||h|| \\ r = v'd \\ A = \begin{pmatrix} \Lambda_{t-1} & 0 \\ 0 & 0 \end{pmatrix} \frac{N_{t-1}}{N_t} \\ \quad + \begin{pmatrix} gg' & gr' \\ rg' & rr' \end{pmatrix} \frac{N_{t-1}}{N_t^2} \end{cases}$
$\Lambda_t = eigenvalue(A)$
$R = eigenvector(A)$
$U_t = [U_{t-1}\, v]R$
return $(\bar{x}_t, U_t, \Lambda_t, N_t)$
```

Algorithm 4.2: INCRE_FORGET($\bar{x}_{t-1}, U_{t-1}, \Lambda_{t-1}, N_{t-1}, z$)

```
$N_t = N_{t-1} - 1$
$\bar{x}_t = \frac{\bar{x}_{t-1} N_{t-1} - z}{N_t}$
$d = \bar{x}_{t-1} - z$
$g = U'_{t-1} d$
$A = \Lambda_{t-1} \frac{N_{t-1}}{N_t} - \frac{gg'}{N_{t-1}}$
$\Lambda_t = eigenvalue(A)$
$R = eigenvector(A)$
$U_t = U_{t-1} R$
Prune the subspace bases in $U_t$with eigenvalue too
  small in $\Lambda_t$
return $(\bar{x}_t, U_t, \Lambda_t, N_t)$
```

5. EXPERIMENTS ON THE CHIL CORPUS

For experiments, we use the CHIL seminar database collected at the University of Karlsruhe (UKA), Germany. The set contains 12 seminars recorded in the UKA smart room (see Fig. 1), with seven seminars collected during 2003 (referred to as the Sem03 data), and five recorded during 2004 (Sem04 data). For every seminar, segments are allocated to the development and test sets, each containing approximately 9,000 frames for each of the four available camera views per seminar, and labeled with 2D and 3D face / head location information. The proposed algorithm is evaluated using the following metrics:

3D error: Mean Euclidean 3D distance in millimeters between the estimated and the ground truth position of the head center in 3D coordinates. In addition, "% err $<$ 300" is the percentage of time instants, where the 3D error is smaller than 300 mm.

2D error: Mean of Euclidean 2D distance in millimeters between the projection on the smart room floor of the estimated 3D head center and that of the corresponding ground truth projection. Furthermore, "% err $<$ 300" is the percentage of time instants, where the 2D error is smaller than 300 mm.

For comparison purposes, a number of alternative tracking systems are considered on the same data:

BGS: It is based on the IBM "PeopleVision" system [15], properly modified for use on the CHIL data. The system uses background

Table 1. Comparison of 3D head-tracking performance of various algorithms on the CHIL seminar task.

Data	Sem03				
Metrics	Proposed	SWF	BGS	MS	NC
3D error (mm)	140.0	270.2	278.4	253.9	1649.4
3D err $<$ 300	92.9%	82.5%	81.2%	84.6%	13.2%
2D err (mm)	123.6	217.3	204.7	228.3	1230.7
2D err $<$ 300	93.25%	84.3%	84.1%	85.3%	14.6%
Data	Sem04				
Metrics	Proposed	SWF	BGS	MS	NC
3D error (mm)	155.2	267.4	480.3	467.4	1852.4
3D err $<$ 300	95.4%	83.6%	47.7%	78.9%	10.9%
2D err (mm)	141.8	208.9	436.9	441.1	1635.1
2D err $<$ 300	95.6%	85.7%	57.1%	80.7%	12.6%

subtraction based object detection that utilizes a multiple Gaussian color model at each pixel and object tracking based on the tracking method described in [16].

MS: This is our prior system, presented in [10]. It uses mean-shift color based tracking and a simpler initialization mechanism: Motion history to segment the foreground region, multi-view face detector applied around the foreground region, and calibration information to obtain the final estimate. Tracking drift detection is similar to the proposed system.

SWF: It is close to the proposed system, but with no forgetting mechanism applied in the adaptive subspace based tracking. It provides a means to demonstrate the value of the forgetting mechanism, introduced in this paper.

NC: It is identical to the system proposed in this paper, however with no drift detection mechanism present. It is provided as a means to demonstrate the value of drift detection.

As it becomes clear from the table, the proposed system significantly outperforms all others. This is also demonstrated in Fig. 4, where the benefit of the forgetting mechanism is clearly shown.

6. SUMMARY

In this paper, we proposed a novel vision system that reduces tracking drift of adaptive, subspace based appearance models by introducing a "forgetting mechanism" in the subspace update process. For initializion, a dynamic programming approach was used to analyze raw face detection results in both temporal and spatial domains. The system was evaluated on the CHIL seminar database, where it dramatically reduced tracking error over previously tried techniques, achieving an average of about 15 cm in 3D head tracking error.

7. REFERENCES

[1] M. Isard and A. Blake, "Contour tracking by stochastic propagation of conditional density," in *Proc. Europ. Conf. Computer Vision*, pp. 343–356, 1996.

[2] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *Proc. Int. Conf. Computer Vision Pattern Recog.*, vol. 2, pp. 142–149, 2000.

[3] H. Tao, H.S. Sawhney, and R. Kumar, "Dynamic layer representation with applications to tracking," in *Proc. Int. Conf. Computer Vision Pattern Recog.*, vol. 2, pp. 134–141, 2000.

[4] M.J. Black and A. Jepson, "Eigentracking: Robust matching and tracking of articulated objects using a view-based representation," *Int. J. Computer Vision*, 26(1): 63–84, 1998.

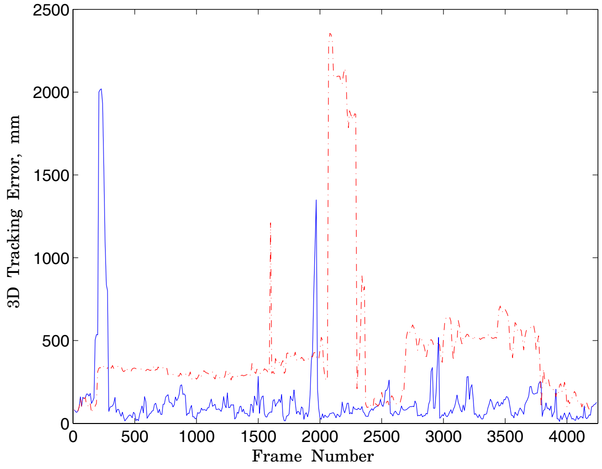

Fig. 4. Typical tracking behavior of the proposed system (solid line), compared with its variant (SWF: dashed line) with no forgetting mechanism, evaluated over a CHIL seminar segment.

[5] A.D. Jepson, D.J. Fleet and T.F. El-Maraghi, "Robust online appearance models for visual tracking," *IEEE Trans. Pattern Anal. Machine Intell.*, 25(10): 1296–1311, 2003.

[6] R.T. Collins, Y. Liu and M. Leordeanu, "Online selection of discriminative tracking features," *IEEE Trans. Pattern Anal. Machine Intell.*, 27(10): 1631–1643, 2005.

[7] B. Han and L. Davis, "On-line density-based appearance modeling for object tracking," in *Proc. Int. Conf. Computer Vision*, Beijing, 2005.

[8] J. Lim, D. Ross, R.-S. Lin, and M.-H. Yang, "Incremental learning for visual tracking," in *Proc. NIPS*, 2004.

[9] CHIL: "Computers in the Human Interaction Loop," `http://chil.server.de`

[10] Z. Zhang, G. Potamianos, A. Senior, S. Chu, and T. Huang, "A joint system for person tracking and face detection", in *Proc. Int. Wksp. Human-Computer Interaction*, Beijing, China, 2005.

[11] Z. Zhang, G. Potamianos, M. Liu, and T.S. Huang, "Robust multi-view multi-camera face detection inside smart rooms using spatio-temporal dynamic programming," *Proc. Int. Conf. Automatic Face Gesture Recog.*, Southampton, United Kingdom, 2006.

[12] S.Z. Li and Z. Zhang, "FloatBoost learning and statistical face detection," *IEEE Trans. Pattern Anal. Machine Intell.*, 26(9): 1112–1123, 2004.

[13] J. Ho, K.-C. Lee, M.-H. Yang, and D. Kriegman, "Visual tracking using learned linear subspaces," in *Proc. Int. Conf. Computer Vision Pattern Recog.*, vol. 1, pp. 782–789, 2004.

[14] P. Hall, D. Marshall, and R. Martin, "Merging and splitting eigenspace models," *IEEE Trans. Pattern Anal. Machine Intell.*, 22(9): 1042–1049, 2000.

[15] A. Hampapur, S. Pankanti, A.W. Senior, Y.-L. Tian, L. Brown, and R. Bolle, "Face cataloger: Multi-scale imaging for relating identity to location," in *Proc. IEEE Conf. Advanced Video Signal Based Surveillance*, pp. 13–20, 2003.

[16] A. Senior, "Tracking with probabilistic appearance models," in *Proc. Int. Work. on Performance Evaluation of Tracking and Surveillance Systems*, 2002.