SIViP (2007) 1:163–178
DOI 10.1007/s11760-007-0018-3

ORIGINAL PAPER

Joint face and head tracking inside multi-camera smart rooms

Zhenqiu Zhang · Gerasimos Potamianos · Andrew W. Senior · Thomas S. Huang

Received: 26 October 2006 / Revised: 26 March 2007 / Accepted: 30 March 2007 / Published online: 30 May 2007

Abstract The paper introduces a novel detection and tracking system that provides both frame-view and world-coordinate human location information, based on video from multiple synchronized and calibrated cameras with overlapping fields of view. The system is developed and evaluated for the specific scenario of a seminar lecturer presenting in front of an audience inside a "smart room", its aim being to track the lecturer's head centroid in the three-dimensional (3D) space and also yield two-dimensional (2D) face information in the available camera views. The proposed approach is primarily based on a statistical appearance model of human faces by means of well-known AdaBoost-like face detectors, extended to address the head pose variation observed in the smart room scenario of interest. The appearance module is complemented by two novel components and assisted by a simple tracking drift detection mechanism. The first component of interest is the initialization module, which employs a spatio-temporal dynamic programming approach with appropriate penalty functions to obtain optimal 3D location hypotheses. The second is an adaptive subspace learning based 2D tracking scheme with a novel forgetting mechanism, introduced to reduce tracking drift and increase robustness. System performance is benchmarked on an extensive database of realistic human interaction in the lecture smart room scenario, collected as part of the European integrated project "CHIL". The system consistently achieves excellent tracking precision, with a 3D mean tracking error of less than 16 cm, and is demonstrated to outperform four alternative tracking schemes. Furthermore, the proposed system performs relatively well in detecting frontal and near-frontal faces in the available frame views.

Keywords Person tracking · Face detection · Multi-camera tracking · Dynamic programming · Adaptive subspace tracking · Mean-shift tracking · AdaBoost · Lecture data · Smart rooms

1 Introduction

Visual detection and tracking of humans is an important problem with numerous applications that range from automated surveillance to interfaces for human–computer interaction. In general, robust human tracking in complex scenes is challenging. In some circumstances however, multiple time-synchronous and calibrated camera sensors with overlapping fields of view may be available, from which both frame-view and world-coordinate human location information can be derived. In such scenarios, efficiently combining frame-level appearance-based human detection with temporal and spatial constraints constitutes a viable approach. This can simultaneously provide both desired types of location information with improved accuracy, while avoiding reliance on any form of background modeling or motion estimation.

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.

Z. Zhang · T. S. Huang
Beckman Institute,
University of Illinois, Urbana, IL 61801, USA
e-mail: zzhang6@uiuc.edu

T. S. Huang
e-mail: huang@ifp.uiuc.edu

G. Potamianos (✉) · A. W. Senior
IBM T.J. Watson Research Center,
Yorktown Heights, NY 10598, USA
e-mail: gpotam@us.ibm.com

A. W. Senior
e-mail: aws@us.ibm.com

The paper introduces a novel human tracking vision system employing these principles, developed and evaluated for the specific scenario of tracking a seminar lecturer presenting inside a "smart room" in front of an audience.

This scenario is of central focus in the European integrated project Computers in the Human Interaction Loop (CHIL [1]). In CHIL, smart rooms have been set up, equipped with multiple audio and visual sensors that include a minimum of four calibrated and time-synchronous cameras with highly overlapping fields of view, located at the room corners (see also Figs. 1 and 2). Numerous seminars have been recorded in such rooms providing a large multi-sensory and multimodal database of real human interaction [2]. The resulting CHIL corpus, annotated with a wealth of multimodal information, has been crucial to the development and evaluation of technologies for perception of humans in the lecture scenario of interest [3,4]. Prominent among such technologies is the task of locating the lecturer's head position, both in the three-dimensional (3D) space—in the form of head centroid world coordinates, as well as in the available two-dimensional (2D) frame views—as bounding boxes of visible faces [5]. Such location information can be further utilized in support of numerous audio-visual perception technologies: For example, 2D face information is useful for person identification [6], whereas 3D location coordinates can be employed in acoustic beamforming for far-field automatic speech recognition [7], as well as to obtain close-up presenter views based on steerable pan-tilt-zoom cameras or camera selection schemes [8,9]. The views can further assist identification [10] and audio-visual speech technologies [11], among others, with obvious utility in lecture indexing and understanding of the interaction.

It becomes clear that for the CHIL lecture scenario described above a visual system that combines face detection, tracking, and multi-camera processing is both feasible and desirable. This paper introduces such a system, developed to provide both 2D-face and 3D-head location information of a single person (the lecturer) in CHIL seminars. Like most 3D approaches, the proposed algorithm consists of a sequence of 3D initialization and tracking phases, with a tracking drift

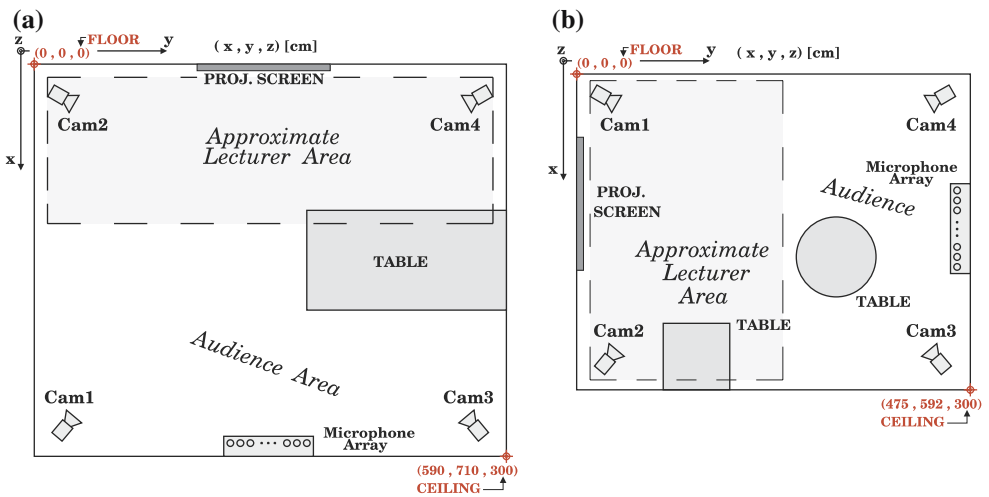

Fig. 1 Overview of the CHIL lecturer video tracking task. Schematic diagrams of the smart rooms located at two CHIL project partners: **a** Universität Karlsruhe (UKA), Germany, and **b** Istituto Trentino di Cultura (ITC), Italy. The CHIL lecture corpus, used in our experiments for single-person (lecturer) tracking, has been collected at these two sites (see also Fig. 2)

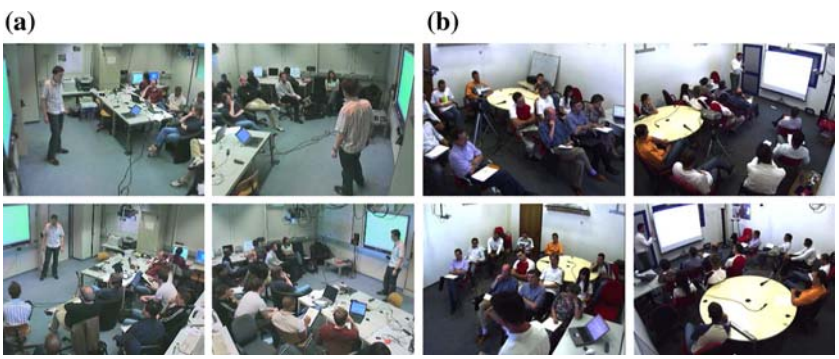

Fig. 2 Examples of synchronous four camera views of the **a** UKA and **b** ITC data, part of the CHIL lecture corpus [2]. In such recordings, a standing subject presents a lecture in front of a small (mostly sitting) audience. Notice the highly overlapping fields of view of the four cameras, set up to ensure that at least two cameras capture the lecturer head at any given instant

detection mechanism controlling the switch between the two. Similarly to other works, all its stages depend on 2D information from separate views to obtain 3D location world coordinates based on camera calibration [12].

However, the proposed system deviates from other research efforts that focus on the 2D or 3D tracking problems alone, in that it jointly considers them within a single framework, in order to improve both 2D-face and 3D-head localization accuracy. This is accomplished by relying heavily on the appearance model of the tracked object—here the lecturer's head, as viewed in 2D by the available cameras. For this purpose, off-the-shelf statistical classifiers of human faces are utilized, in particular AdaBoost-like face detectors, appropriately extended to address the head pose variation observed in the smart room scenario. As a result, in the developed system, 2D face detection plays a pivotal role in 3D head tracking, being employed in system initialization and in detecting possible tracking drift. Similarly, 3D tracking determines the 2D frame regions where a face detector may be subsequently applied. An additional differentiator of the proposed system is that no form of motion estimation or background modeling is needed. The algorithm therefore remains robust to the unpredictability of motion, occlusion, and background changes in the heavily cluttered CHIL smart rooms. This is in contrast to all alternative tracking systems in the literature (to our knowledge) that address the smart room scenario of interest [13–20].

Two additional components in the proposed system complement the appearance module, implementing a number of novel ideas: One is the initialization module that employs a spatio-temporal dynamic programming approach to obtain optimal 3D location hypotheses. For this purpose, while scoring candidate hypotheses, the adopted implementation penalizes not only large trajectory discontinuities over time, but also accounts for hypothesis appearance similarity between camera views. The second component of interest is a 2D tracking module, used as part of the 3D tracking phase. This component utilizes an adaptive subspace learning based scheme [21]. A novel forgetting mechanism is introduced into this technique to reduce tracking drift and increase robustness to illumination and head pose variation. Furthermore, tracking is applied on only two of the four available camera views, selected based on the initialization component. This results in considerable speed-up.

Finally, the extensive benchmarking of the proposed approach constitutes an important aspect of the paper, breaking away from the toy-problem or small-scale evaluation paradigm that often accompanies other works in the area. In particular, the developed system is benchmarked on all three parts of the CHIL lecture corpus [2]. This is a large database that exhibits significant data variability, with no artificially imposed constraints in the human interaction and behavior patterns, thus allowing meaningful technology development, evaluation, and algorithmic comparisons [5]. Furthermore, the proposed system is compared to a number of 3D tracking methods, ranging from small algorithmic variations of it to significantly different approaches [20,22].

The rest of the paper is organized as follows: Section 2 briefly discusses literature work relevant to this paper. Section 3 provides an in-depth presentation of the proposed system and its components. Section 4 describes alternative systems considered in our experiments on CHIL lecture data, which are subsequently presented in Sect. 5. Finally, a brief summary and discussion in Sect. 6 conclude the paper.

2 Related Work

Much work has been devoted to the core problems of human detection and tracking that constitute the focus of this paper. For this purpose, human body models are often used, ranging from simplistic blob appearance or cylindrical shape models [23] to more complex articulated ones [24]. An alternative approach to these problems is detecting and tracking human faces.

For face detection, machine learning based techniques are widely considered as the most effective, for example based on neural networks [25], support vector machines [26], network of linear units [27], or the AdaBoost approach [28]. Alternative methods using traditional image processing algorithms based on color and edge information [29], or optimization to match learned shape and/or appearance to data [30] have also been shown to achieve good performance. Many such techniques can be further extended to handle detecting faces under varying head pose; for example [31], where pose-based appearance frameworks are proposed, or the multi-pose face detection work of Li et al. [32], where "FloatBoost", an AdaBoost variant, is employed. The latter approach is used in our proposed system.

Similarly, for tracking faces, various target representations have been used in the literature, such as parameterized shapes [33], color distributions [34], image templates [35] and the eigenspace approach [36]. 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 online, for example the EM-algorithm based technique of [37], the feature selection mechanism of [38], and the parametric statistical appearance modeling technique in [39]. An interesting non-parametric approach appears in Lim et al. [21], where the appearance subspace is learned online by an efficient sequential algorithm for principal component analysis (PCA), updated with the incoming data vectors. An extension of this technique is employed in our proposed system.

In general however, real interaction scenarios, such as in the CHIL domain, present significant challenges to most face detection and tracking algorithms, for example partially occluded and low-resolution faces, as well as lighting and head-pose variations. These difficulties can often be successfully addressed, only if additional information is available in the form of multi-camera input, in order to reduce spatial uncertainty in the scene [40]. Naturally, some researchers have begun to exploit multiple camera views where they are available, and several tracking systems attempt to fuse information from the available sensors to yield 3D tracking results [10,41–43], using for example Kalman filters [44], particle filters [45], or just scene and camera geometry [40].

The above ideas have found their way into a number of papers for tracking the lecturer in the CHIL scenario. In 2D face tracking work reported in [13,14], statistical face detection is assisted by either a motion model or a combination of foreground-background segmentation [46] and 2D Kalman filtering. However, neither system utilizes 3D information. A few other works aim to provide 3D head information in the CHIL scenario of interest [15–20]. These differ from our proposed system in various aspects, most importantly that they do not focus directly on the 2D face appearance information (with the exception of [20]), but rather model and track larger parts of the human body. For this purpose, they all use background modeling [15–17] or motion information [18–20]. The extracted camera view information is then combined across views by employing either triangulation-based, decision fusion mechanisms [17,19,20], or likelihood fusion by means of particle filters [16,18]. An alternative technique appears in [47], where histogram features are directly combined across camera views within a 3D kernel based tracking framework—a process akin to feature fusion. That system however lacks an initialization component.

3 The proposed tracking system

We now proceed to describe the developed tracking system. As already discussed in the Introduction, this constitutes a joint face- and head-tracking approach, developed to yield both the 3D head centroid location of the lecturer inside the CHIL smart room, as well as the 2D bounding boxes of visible lecturer faces (ranging from frontal to profile) in the available camera views. We first briefly present an overview of the entire system, followed by a detailed discussion of its main algorithmic components.

The following notation will be used in this section: $H^{(t)}$ will denote a hypothesis at instant t concerning the 3D world coordinates (x_t, y_t, z_t) of the lecturer's head centroid. Similarly, $h_c^{(t)}$ will represent a hypothesized "visible face"

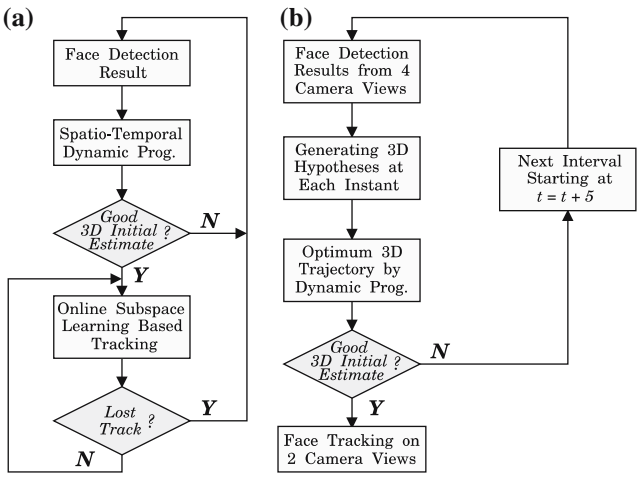

Fig. 3 Block diagram of the developed multi-camera 3D head tracking system. **a** Overview; **b** Initialization

at instant t in camera view $c \in \mathcal{C}$, where $\mathcal{C}$ is the set of available cameras (here, four). Face hypothesis h contains 2D information about the face bounding box, $(u, v, \Delta u, \Delta v)$, namely 2D center coordinates, box height and width. The collection of pixels within h will be denoted by $\mathbf{h}$.

3.1 System overview

The overview diagram of the 3D head tracking system is given in Fig. 3a. It basically consists of an initialization and a tracking component, with tracking drift detection controlling the switch between these two modes. For its initialization, multi-pose face detectors are first applied to all four camera views in the smart room (also referred to in this work as a "quad-frame"—see Fig. 2). Details are provided in Sect. 3.5. Subsequently, spatio-temporal information of the face detection results over ten consecutive quad-frames is integrated within a dynamic programming framework, to provide robust initialization. Details are described in Sect. 3.2 (see also Fig. 3b). Following initialization, a 2D tracking component kicks in, operating independently in two only camera views selected based on the initialization step. Details of the tracking algorithm, which is based on online adaptive subspace learning, are presented in Sect. 3.3. Finally, an important aspect of the system is the re-initialization decision. This is described in Sect. 3.4.

In addition to 3D head tracking, the developed system performs 2D face localization, based on the 3D result. Such result provides the approximate region within the 2D frame views, where a visible face could be present, in the following manner: As mentioned above, the 3D system uses 2D subspace tracking on two only camera views. For these views, the expected face location is therefore immediately available.

For the remaining two camera views, the system considers the projection of the 3D head position estimate (using camera calibration) to obtain an estimate of the head's 2D location in the image frames. Following this step, multi-pose face detection (see Sect. 3.5) is applied around the estimated head center in each camera view. If the face detector locates a face, this is accepted. If there is no face detection result, then one of the following two cases occurs: (a) If the camera view in question is one of the two views that have been used in tracking at that instant, the raw 2D tracking result (i.e., the tracked face box) is returned as the face detection output. (b) If however the camera is not a 2D tracking view, no face output is produced.

3.2 Spatio-temporal 3D initialization

Robust initialization is a crucial component in every tracking scheme. In the proposed system, initialization is driven by the face detection module described in detail in Sect. 3.5. In particular, trained AdaBoost-like multi-pose face detectors are applied on all four camera views (over the entire quad-frame) and over all time instants during the initialization phase. However, the resulting detected faces prove insufficient to lead to robust 3D initialization by triangulation alone [40]. This is due to high rates of false positives and missed faces, as discussed in Sect. 3.5 and quantified in the experiments (Sect. 5.4)—see also Fig. 4.

Given the challenging nature of face detection in the CHIL scenario, the developed system seeks to utilize additional information, in the form of temporal (video sequences) and spatial (multiple camera views) context. The resulting algorithm integrates both temporal and spatial information from frame-level face detection results into a *dynamic programming* (DP) framework, schematically depicted in Fig. 3b. In summary, following face detection, 3D hypotheses of the presenter's head location are generated using the calibration information, based on the spatial consistency of the detection result from different camera views. Then, DP applied on the results over ten consecutive quad-frames is used to search for the optimal trajectory of the presenter's head centroid in the 3D space, based on appropriately defined penalty functions. If the optimal trajectory is accepted compared to a threshold, the result is fed into the tracking component described in Sect. 3.3; otherwise the process is iterated with a five frame shift until an acceptable trajectory is determined. An example of the proposed spatio-temporal initialization scheme applied on CHIL lecture data is depicted in Fig. 4. Details of the implementation follow.

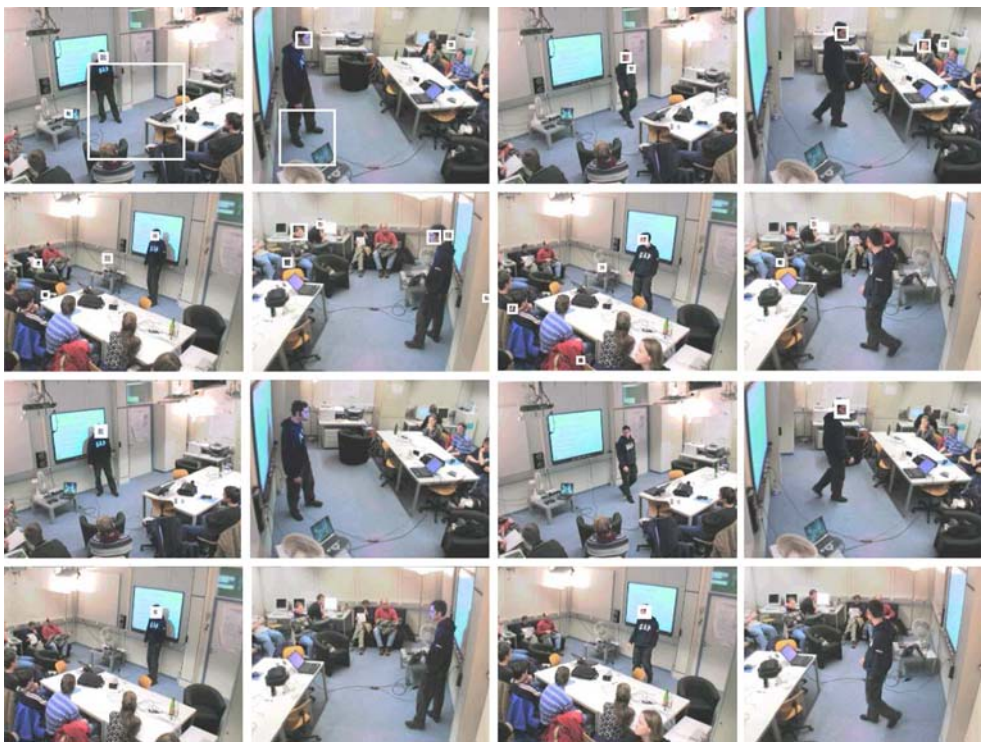

Fig. 4 Spatio-temporal face detection depicted at two instants, for all four camera views. *Upper-row:* Based on frame-level FloatBoost face detection, with no spatial and temporal information utilized. *Lower-row:* After the proposed dynamic programming. Notice that in the latter case, single faces (frontal or profile) are only depicted for the two selected camera views that correspond to the optimal hypothesis

3.2.1 Generating 3D hypotheses

Assuming n_i face detections per camera view, there could be up to

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

candidate 3D head locations at each time instant, obtained via pair-wise triangulation of detected face bounding box centers, using for example the direct linear transformation (DLT) method [40]. A few of these hypotheses can be readily rejected, for example when large inter-ray distances of the 2D-to-3D maps are observed, or based on collection-site specific spatial constraints. The latter can be learned from development data, and are imposed to distinguish the lecturer from audience members (see also Fig. 1). These constraints result in about half of the room floor surface being allowable for the presenter's (x,y) location, whereas a 400 mm height range (1,500–1,900 mm) is imposed on the z-axis location coordinate. As a result of this process, multiple 3D hypotheses

$$H_i^{(t)} = \text{DLT}\,(h_{k_i}^{(t)}, h_{l_i}^{(t)}) \tag{1}$$

are generated at every time instant t, where indexes k_i, l_i specify the face hypotheses in two camera views that yield $H_i^{(t)}$. Hence in this framework, each $H_i^{(t)}$ contains not only the 3D location coordinates of the hypothesized head centroid, but also indexing information about the two camera views that generated it.

3.2.2 Trajectories of 3D hypotheses

Following generation of a pool of 3D head centroid hypotheses at each time instant t, the next step is to perform dynamic programming over the temporal window of interest, in order to obtain the optimal temporal sequence (path or trajectory) of 3D location hypotheses. For this purpose, two main 3D-path cost components are employed. One is a traditional transition cost that penalizes path discontinuities over time. An additional local cost complements it, based on a similarity measure of the 3D hypothesis. This is introduced to reward consistency among the face detection results that generated the hypothesis via (1). As a result, a path

$$\mathbf{H} = \left\{ H_{i_1}^{(t_1)}, H_{i_2}^{(t_2)}, \ldots, H_{i_n}^{(t_n)} \right\} \tag{2}$$

based on n 3D hypotheses of head centroids at times $t_1 < t_2 < \cdots < t_n$ has a trajectory cost associated to it, given by

$$C^{(t)}(\mathbf{H}) = t_1\, C_B \;+\; (t - t_n)\, C_E \;+\; \sum_{k=1}^{n-1}(t_{k+1} - t_k)\, C_I + \sum_{k=1}^{n-1} C_T(H_{i_{k+1}}^{(t_{k+1})} | H_{i_k}^{(t_k)}) \;+\; \sum_{k=1}^{n} C_L(H_{i_k}^{(t_k)}), \tag{3}$$

over time interval $[\,0\,,\,t\,]$, where $t \geq t_n$. In (3), $C_T(\bullet|\bullet)$ and $C_L(\bullet)$ denote the transition and local similarity costs, respectively. In addition to those, three constant costs are introduced to account for missing 3D hypotheses, or to allow skipping unreliable ones (by essentially duplicating a prior hypothesis) in some of the instants over the temporal window of interest. The three costs, C_B, C_I, C_E are used for this purpose at the beginning, intermediate, or ending part of the trajectory, respectively. Additional details of the components in (3), as well as the hypothesis search follow.

3.2.3 Local similarity cost

This is used to evaluate the hypothesis at the current instant on the basis of the available camera views that generated it via (1), exploiting spatial information by means of local appearance. The assumption is that if the candidate hypothesis corresponds to an actual 3D object, then the corresponding face regions in the two camera views should have similar color histograms. The cost computation is based on the Bhattacharyya coefficient, and is defined as (see also (1))

$$C_L(H_i^{(t)}) \;=\; -\,\alpha \sum_{b=1}^{m} \sqrt{p_b(\mathbf{h}_{k_i}^{(t)})\; p_b(\mathbf{h}_{l_i}^{(t)})}\,, \tag{4}$$

where $\{p_b(\mathbf{h}) : b = 1, \ldots, m\}$ denotes the m-bin color histogram, based on the face candidate pixel values $\mathbf{h}$, and α is a scalar value used in order to balance the contributions of (4) and (5) in (3).

In our implementation, p is taken to be the 30-bin histogram of the H component of the color HSV space. Furthermore, and in order to improve robustness, the face candidate regions in the computation of (4) are extended: Histograms are computed over rectangles taken to be approximately double (in height only) the detected face bounding boxes $h_{k_i}^{(t)}$ and $h_{l_i}^{(t)}$.

3.2.4 Transition cost

The transition cost exploits temporal information, and it is used to penalize non-smooth trajectories, based on the 3D

distance between temporally consecutive hypotheses. The cost is specified using Gaussian diffusion, computed between 3D hypotheses $H_i^{(t)}$ and $H_j^{(t-1)}$, as

$$C_T\left(H_i^{(t)}|H_j^{(t-1)}\right) = \frac{1}{2}\log|\Sigma| + \frac{3}{2}\log 2\pi + \left(H_i^{(t)} - H_j^{(t-1)}\right)^T \times \Sigma^{-1}\left(H_i^{(t)} - H_j^{(t-1)}\right). \quad (5)$$

In our system, the covariance matrix Σ is set to diagonal matrix (100,100,100), assuming that 3D hypothesis coordinates are in mm.

3.2.5 Hypothesis search

The searching scheme employs the standard dynamic programming approach, based on cost equation (3) — but with a few twists to better adapt to the task at hand. Available at a given instant t are a pool of local hypotheses $H_i^{(t)}$, $i = 1,\ldots,m$, and the active trajectories up to $t-1$, which we denote by $\mathbf{H}_j^{(t-1)}$, $j = 1,\ldots,n$, extending the notation in (2). The latter are accompanied by scores $g_j^{(t-1)}$ that specify the trajectory cost up to $t-1$, based on (3). Then, the active hypotheses at t are obtained as $\mathbf{H}_i^{(t)} = \{\mathbf{H}_{\hat{\jmath}(i)}^{(t-1)}, H_i^{(t)}\}$, where

$$\hat{\jmath}(i) = \underset{j=1,\ldots,n}{\operatorname{argmin}}\{\, g_j^{(t-1)} + C_T(H_i^{(t)}|\, H_j^{(t-1)}) + C_L(H_i^{(t)})\,\},$$

with the new score $g_i^{(t)}$ being the optimal value of the above minimized expression. In addition to the updated trajectories, active hypotheses $\mathbf{H}_j^{(t-1)}$ may remain "alive" as $\mathbf{H}_j^{(t)} = \{\mathbf{H}_j^{(t-1)}, H_j^{(t-1)}\}$ (slight notation abuse) with a constant penalty C_I added to their score (see (3)). To speed up computations, pruning is performed among the resulting pool of paths, by allowing at most six trajectories to be kept active at any instant t. Furthermore, the scheme is terminated at the 10th quad video frame ($t_{\text{end}} = t_{\text{init}} + 10$). The global optimal trajectory is then obtained by choosing the active hypothesis with the minimum score at $t = t_{\text{end}}$.

In addition, a maximum acceptable score is defined, providing a mechanism to reject the final hypothesis (and hence trigger a new search), if its total cost exceeds a fixed threshold. This threshold, as well as parameters $C_I = C_B = C_E$ and α in (3) and (4), are tweaked empirically, based on detection and false alarm rates on CHIL development data. In the case that the optimal trajectory is rejected, a five quad-frame shift is applied, and the search gets re-initialized. The returned optimal trajectory defines the two camera views on which 2D tracking is to commence, as discussed next.

3.3 Adaptive subspace 2D tracking

Following successful initialization, a 3D hypothesis is obtained as the last element of the optimal (minimum score) spatio-temporal path at time instant $t_o \doteq t_{\text{end}}$. This hypothesis, denoted by

$$H^{(t_o)} = \text{DLT}\,(h_{c'}^{(t_o)}, h_{c''}^{(t_o)}),$$

contains the two face detection results and the indexing information of the two camera views, $c', c'' \in \mathcal{C}$, that generated it. Such information allows the tracking phase of the algorithm to commence. This stage consists of two separate 2D tracking processes, running independently and in parallel for each of these two camera views. The 2D processes are based on an adaptive PCA subspace approach that tracks the face bounding box within the single-camera frame sequence. Therefore, at each time instant $t > t_o$, the two trackers generate face bounding boxes $h_c^{(t)}$, $c \in \{c', c''\}$. The 3D head centroid location can then be easily obtained via triangulation as $H^{(t)} = \text{DLT}\,(h_{c'}^{(t)}, h_{c''}^{(t)})$, assuming that no tracking drift is detected.

The motivation behind this scheme is to reduce computations by tracking using the bare minimum of camera views (two), sufficient for 3D triangulation, but also to do so in the specific views where visible faces (frontal or profile) are expected. Such views contain more discriminating information, as opposed to views that capture the back of the lecturer's head. In addition, they enable the verification of whether the hypothesized tracked object is indeed a visible face, by applying a face detector in its region. This is crucial in detecting possible tracking problems (see Sect. 3.4). Furthermore, the 2D tracking results may readily provide desired 2D face information in the camera views in question, as discussed in Sect. 3.1.

At the heart of the proposed scheme lies the 2D PCA subspace tracking approach. As discussed in Sects. 1 and 2, adaptability of the subspace to the observed conditions is crucial in improving tracking robustness in the dynamic CHIL scenario, mainly due to variations in head-pose and lighting. Such approaches have already been proposed in the literature, for example in [21]. There, 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 objects over long durations, since the noise introduced during tracking eventually could bias the PCA subspace away from

the characteristic appearance of the desired tracked object. In [48], an L_∞ norm subspace is fitted to the past frames incrementally by Gramm-Schmitt orthogonalization. Though the subspace with L_∞ norm has the advantage of timely incorporating observation novelties into the subspace representation [48], it runs the risk of tracking drift due to its lack of robustness to noise and outliers. PCA on the other hand offers freedom to perform dimensionality reduction and thus ignore tracking noise and assist outlier rejection based on reconstruction error [36]. Therefore, the proposed system adopts the incremental PCA subspace learning approach. In particular, Hall's mechanism [49] is employed to incrementally update the PCA subspace given new observations. In addition, our proposed system also allows subspace adjustment, by eliminating distant past observations in the subspace. This introduces a forgetting mechanism that is absent in Lim's approach [21].

The proposed 2D adaptive subspace tracking scheme consists of three steps, at each time instant (frame) t, as discussed next. The presentation refers to faces, but of course the scheme is more general.

(a) *Localization*: The first step is to estimate the new face location at instant t, $h^{(t)}$, based on the prior face location, $h^{(t-1)}$, and the available PCA subspace of face appearance at t–1 (for simplicity, we drop the camera index in the notation). Let us denote the current PCA subspace by ($\bar{\mathbf{h}}^{(t-1)}$, $U^{(t-1)}$, $\Lambda^{(t-1)}$, $N^{(t-1)}$), with its elements representing, respectively, the mean vector of face appearances, the matrices of retained eigenvectors and eigenvalues, and the current number of observations modeled. The new face location at t is then obtained as

$$h^{(t)} = \underset{h \in \mathcal{N}(h^{(t-1)})}{\arg\min} \| (\mathbf{h} - \bar{\mathbf{h}}^{(t-1)}) - U^{(t-1)} U^{(t-1)\,T} (\mathbf{h} - \bar{\mathbf{h}}^{(t-1)}) \|_2, \quad (6)$$

where the minimization occurs over a set of candidate face bounding boxes in the "neighborhood" $\mathcal{N}(h^{(t-1)})$ of the previous face. Note that in (6), the minimized functional corresponds to the distance from the PCA space of the vectors of candidate face pixels, $\mathbf{h}$, within the corresponding face bounding boxes h.

(b) *New sample inclusion into subspace*: Once the new face "observation" $h^{(t)}$ becomes available, its pixel values vector $\mathbf{h}^{(t)}$ gets recruited into the PCA subspace. The subspace can be adapted in an incremental fashion, as described in Alg. 1 of Fig. 5, thus avoiding recomputing the subspace from all its samples.

(c) *Old sample exclusion from subspace*: Following inclusion of the new observation, the PCA subspace receives a second update by excluding a past distant observation vector $\mathbf{h}^{(t-m)}$. This forgetting mechanism is performed as described in Alg. 2 of Fig. 5, avoiding recalculation of the entire subspace. Notice that in contrast to step (b), the process occurs only once the subspace reaches its "steady state" of containing $N^{(t)} = m$ samples, or equivalently for $t \geq t_o + m$.

In our particular implementation, the proposed system employs the most recent $m = 50$ frame observations to construct the PCA subspace. Hence, following tracking initialization, the forgetting mechanism does not commence until after 50 frames are observed. For this initial duration, the algorithm remains identical to [21]. The learned subspace has a dimensionality of up to 15, down from a normalized 20×20-pixel data "template" (the un-normalized template size depends on the detected face at the end of the initialization step). Finally, the optimization in (6) occurs over 169 candidate faces of constant size (equal to the detected face size at initialization), with their centers located at equally spaced points within a square four times in size of the initialized face actual size. Tracking therefore occurs in constant scale, with only the face location sought.

Alg. 1: INCLUDE $(\bar{\mathbf{h}}^{(t-1)}, U^{(t-1)}, \Lambda^{(t-1)}, N^{(t-1)}, \mathbf{h}^{(t)})$

$N^{(t)} = N^{(t-1)} + 1$

$\bar{\mathbf{h}}^{(t)} = \dfrac{\bar{\mathbf{h}}^{(t-1)} N^{(t-1)} + \mathbf{h}^{(t)}}{N^{(t)}}$

$d = \bar{\mathbf{h}}^{(t-1)} - \mathbf{h}^{(t)}$

$g = U^{(t-1)\,T} d$

$z = d - Ug$

if $\|z\| = 0$

$A = \Lambda^{(t-1)} \dfrac{N^{(t-1)}}{N^{(t)}} + gg^T \dfrac{N^{(t-1)}}{N^{(t)\,2}}$

else

$v = z/\|z\|$

$r = v^T d$

$A = \begin{pmatrix} \Lambda^{(t-1)} & 0 \\ 0 & 0 \end{pmatrix} \dfrac{N^{(t-1)}}{N^{(t)}} + \begin{pmatrix} gg^T & gr^T \\ rg^T & rr^T \end{pmatrix} \dfrac{N^{(t-1)}}{N^{(t)\,2}}$

$\Lambda^{(t)} = eigenvalue(A)$

$R = eigenvector(A)$

$U^{(t)} = [U^{(t-1)}\, v] R$

return $(\bar{\mathbf{h}}^{(t)}, U^{(t)}, \Lambda^{(t)}, N^{(t)})$

Alg. 2: EXCLUDE $(\bar{\mathbf{h}}^{(t-1)}, U^{(t-1)}, \Lambda^{(t-1)}, N^{(t-1)}, \mathbf{h}^{(t-m)})$

$N^{(t)} = N^{(t-1)} - 1$

$\bar{\mathbf{h}}^{(t)} = \dfrac{\bar{\mathbf{h}}^{(t-1)} N^{(t-1)} - \mathbf{h}^{(t-m)}}{N^{(t)}}$

$d = \bar{\mathbf{h}}^{(t-1)} - \mathbf{h}^{(t-m)}$

$g = U^{(t-1)\,T} d$

$A = \Lambda^{(t-1)} \dfrac{N^{(t-1)}}{N^{(t)}} - \dfrac{gg^T}{N^{(t-1)}}$

$\Lambda^{(t)} = eigenvalue(A)$

$R = eigenvector(A)$

$U^{(t)} = U^{(t-1)} R$

Prune the subspace bases in $U^{(t)}$ with eigenvalue too small in $\Lambda^{(t)}$

return $(\bar{\mathbf{h}}^{(t)}, U^{(t)}, \Lambda^{(t)}, N^{(t)})$

Fig. 5 Brief overview of the incremental adaptive subspace update used for 2D tracking, when including a novel observation (Alg. 1), or excluding a distant past observation (Alg. 2) from the subspace

3.4 Tracking drift detection in 3D

An important aspect of the system is the re-initialization decision, or equivalently, tracking drift detection on basis of the 2D independent tracking results in the two selected camera views. This is based on a combination of local face detection and calibration-based triangulation to test the consistency of the two tracks at the given time. 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, the multi-pose face detectors of Sect. 3.5 are also applied around the two tracking results to determine whether there indeed exists a face object in the local regions of interest (for example, in the proposed system, this is set to a 80×80 pixel region, when running on CHIL seminar data collected at UKA). 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.

3.5 Multi-pose 2D face detection

Face detection is a critical component of the developed system, being used at the initialization (Sect. 3.2) and drift detection stages (Sect. 3.4) of the 3D head tracking sub-system, and in addition being the required step to produce 2D face results, based on the 3D head location estimate, as discussed in Sect. 3.1. Our system adopts a multi-pose face detector approach, with classifiers trained using the FloatBoost technique [32], an AdaBoost variant [28].

3.5.1 AdaBoost and FloatBoost learning

AdaBoost provides a simple yet effective approach for stage-wise learning of a nonlinear classification function [50]. While a good classifier is difficult to obtain at once, AdaBoost learns a sequence of more easily attainable "weak" classifiers, whose performances may be poor, but better than random guessing. It then boosts (combines) them into a "strong" classifier of higher accuracy.

Viola and Jones [28] successfully applied AdaBoost classification to the face detection problem, following earlier work [51]. There, AdaBoost is adapted to solve three issues: (i) Learning effective features from a large feature set; (ii) Constructing weak classifiers, each based on one of the selected features; and (iii) Boosting the weak classifiers into a stronger one. In the particular two-class face detection problem, tens of thousands of simple Haar wavelet-like features are defined, and an appropriate scheme for their selection is designed. The process is carried out sequentially, at each step m selecting a weak classifier $f_m(\mathbf{h})$, simply designed based on its corresponding feature, over the pool of available features. The weak classifier is added into a linear combination of the already chosen weak classifiers in previous steps, resulting to a stronger one, $F_m(\mathbf{h})$. The selection of $f_m(\mathbf{h})$ is based on minimizing the classification error of $F_m(\mathbf{h})$ on an appropriately weighted epoch of the training data. The scheme therefore represents a greedy sequential forward search procedure.

An alternative training algorithm, applied to the face detection problem, appears in [32]. This employs the sequential floating search method [52] that allows feature deletion and controlled backtracking during the strong classifier learning process. In particular, a "conditional exclusion" step is added to AdaBoost training. In it, each of the weak classifiers $f_k(\mathbf{h})$, $0 \leq k \leq m$, that constitute elements of $F_m(\mathbf{h})$ is examined to check whether removing it may reduce classification error of the remaining linear combination. If such situation occurs, and assuming that weak classifier $f_n(\mathbf{h})$ is the one that reduces the error the most when removed, $f_n(\mathbf{h})$ will be deleted, and all classifiers $f_k(\mathbf{h})$, $n < k \leq m$, will be re-learned. The process results in more expensive training compared to traditional AdaBoost, but yields more compact sets of weak classifiers.

Both AdaBoost and FloatBoost learning approaches discussed can be used to combine the successively stronger classifiers into a cascade structure [28,32]. The goal is for the resulting classification structure to quickly reject uninteresting non-face candidates $\mathbf{h}$, while focusing attention to candidates that appear to be face-like (or confused as such). A simple such framework is proposed in [28].

3.5.2 Implementation details

In our implementation, we use the FloatBoost approach [32] to train cascaded (layered) face classifiers using Haar wavelet features [28]. In particular, since faces may be visible in the available camera views with different head poses, we train two detectors, based on clustering visible faces into two groups: Frontal ones that also contain near-frontal faces, and left-side profile ones pooled together with mirrored right-side profile faces. The two face detectors are trained on development set data, on images cropped based on the available CHIL corpus annotations (see also Sect. 5.1). For negative examples (non-faces), training samples are cropped from an image database that does not include faces, as well as non-face regions of CHIL corpus frames. Separate face detectors have been trained for each of the three parts of the CHIL database, discussed in Sect. 5.1. For example, for the "CHIL04" set, 1,606 frontal and 1,542 profile images have been used. Following FloatBoost training, the resulting frontal face detector consists of 15 layers and 576 Haar wavelet features, whereas the profile view one consists of 30 layers and 4,330 features. Notice that during the testing phase, an additional detector of right-side profile view faces is used. This is readily obtained

by mirroring the left-side profile view face detector [32]. An example of detected faces on CHIL data is depicted in the upper rows of Fig. 4.

4 Alternative 3D tracking systems

To evaluate the proposed system, we compare it with a number of alternative 3D tracking approaches in experiments reported in Sect. 5. Two of the systems are only slight variations of the proposed theme; therefore, they are briefly described together with our experiments (see Sect. 5.3). The remaining two however depart significantly from it [20,53], and are overviewed next.

4.1 Motion and mean-shift tracking based system

Similarly to the proposed system, this alternative approach consists of three components, namely 3D initialization, 2D tracking, and drift detection. The latter, as well as the face detection part of the initialization component are identical to the proposed system, as described in Sects. 3.4 and 3.5, respectively. However, the system lacks the more sophisticated spatio-temporal dynamic programming framework for initialization, using instead a motion detection based approach to identify candidate regions for initialization. In addition, it replaces adaptive subspace tracking with the mean-shift tracking algorithm. The two components that differ from the proposed system are briefly discussed below. More information can be found in [20].

4.1.1 Initialization

First, independently for each camera view, motion history is estimated to rapidly determine where movement has occurred. The algorithm used is based on work by Davis and Bobick [54]. Obtaining a foreground silhouette is achieved through subtraction between two consecutive frames instead of background subtraction. As the person moves, the most recent foreground silhouette is copied as the highest value in the so-called "motion history image" (MHI). MHI pixel values that fall below a threshold are set to zero. An example of the algorithm applied to two camera views is depicted in Fig. 6a.

Subsequently, a multi-pose face detector, identical to the one of the proposed system (Sect. 3.5), is applied to the foreground region only (where motion occurred), instead of the whole frame. The detection results for each camera view can then be used to verify whether the detected faces belong to the same person, based on calibration information [40], thus providing the 3D head position. The highest lying 3D position within the general seminar presenter area is returned as the initialization estimate for subsequent tracking.

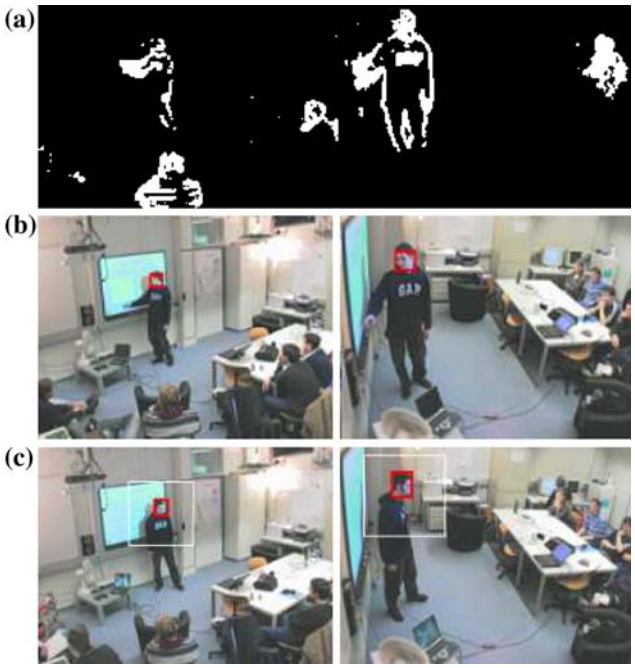

Fig. 6 Examples of processing steps in an alternative 3D head tracking system, based on face detection, motion estimation, and mean shift tracking [20]. **a** Motion history image for two camera views; motion objects are segmented as foreground (*white pixels*). **b** Multi-pose face detection result, after FloatBoost face detectors are applied locally around the resulting foreground region. **c** Local face detection applied within windows around the mean shift based tracking results in the two camera views

The above algorithm could in principle be applied to all four camera views. However, in order to reduce the pool of 3D initialization candidates, two only camera views are being used in the implementation of [20]. These cameras have been selected based on development data from each lecture as the cameras with the highest percentage of (near-)frontal faces. This is possible for "CHIL03" and "CHIL04" data, where development and evaluation sets are available for each of the lectures in the corpus (see Sect. 5.1), and assumes that the lecturer's general location behavior would not change over the duration of the seminar.

4.1.2 Mean shift tracking

Following the initialization component and the successful location of the presenter's face, the algorithm switches into its tracking mode. A color-based face model of the detected face region is first created for tracking in each of the two camera views. In particular, the one-dimensional histogram of the H component in the HSV color space is used for this purpose. The mean shift iteration algorithm is then employed for tracking [34], based on the Bhattacharyya coefficient, around a target position predicted by means of Kalman filtering [44]. The algorithm is applied separately in the two

camera view images to find the best target candidate. Subsequently, triangulation provides the 3D position estimate, with drift detection, as in Sect. 3.4, flagging possible inconsistencies that trigger re-initialization.

4.2 Background subtraction based system

This system constitutes a 3D tracker, developed on top of the IBM "Smart Surveillance Engine" (SSE) 2D tracker [55]. The system employs SSE independently for each of the four available camera views, and then integrates the information in 3D [53].

The 2D component is based around a background-subtraction object detection system [22], which uses a multiple Gaussian color model at each pixel. Objects are tracked in the image plane from frame to frame using the "ColourField" tracking method [55]. A preliminary extension of this system to 3D tracking, called the "Face Cataloger" appears in [10]. There, the 2D tracker was applied independently in two views, and used a head detection algorithm to locate the head center, regardless of pose. The estimated head points from pairs of tracks in the 2D views were then triangulated to determine correspondence and estimate 3D head centroid positions.

Since then, improvements in the underlying 2D tracking algorithm allowed a new 3D tracking algorithm to be developed for the CHIL task [53]. This approach dispensed with the head detector, which had limitations when multiple targets were being tracked, and was found to be unnecessary in lectures, where the head is almost always the highest point of the presenter's body. In this version of the tracker, the underlying improved 2D tracking algorithms of the IBM SSE system are again employed, unmodified from their usual outdoor surveillance configuration. The 2D tracker provides a temporally-smoothed model of the objects observed in each view, together with each object's location, tracked through occlusions. The 2D track information however is not used in the 3D engine; instead, temporal consistency is applied directly in 3D.

In more detail, at each frame, the 2D tracker is applied, and the resulting 2D probabilistic models are used to determine the position of the head top. This is taken to be the point whose y coordinate is the top of the object model bounding box and whose x coordinate is that of the centroid of the upper sixth of the model. The resulting 2D object points are considered as hypotheses for the top of the speaker head, and when coupled with the camera calibration information, each gives a 3D ray, along which the speaker's head might lie. Validation for these hypotheses in other views is then sought, by computing the shortest distance between each pair of such rays from different cameras. All such pairings are evaluated, sorted and compared to a distance threshold of 300 mm, with the closest match considered first. The procedure yields a set of 3D hypothesis points, which can then be associated over time and concatenated into 3D tracks. For this purpose, dynamic programming is employed to find the best track hypothesis through the temporal sequence of 3D head-top hypotheses. The approach uses a beam search with up to N (typically 50) search hypotheses active, to search for the shortest path passing through head location hypotheses. Trajectory costs are given by (3), but with a few differences; namely, $C_T(H_i^{(t)} \mid H_j^{(t-1)}) = \|H_i^{(t)} - H_j^{(t-1)}\|$ and $C_L(H_i^{(t)}) = 0$. At each time instant, all paths are updated, where each path can be retained with no additional evidence (with a penalty), or by adding one of the 3D location hypotheses for that instant. At the end, the lowest cost path is retained as the "best" path through the 3D location hypotheses. Part of this process is depicted in Fig. 7.

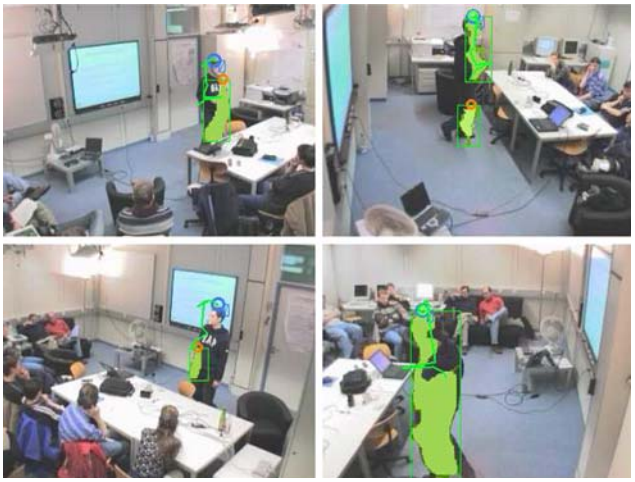

Fig. 7 Detection results for the background subtraction based tracker on four synchronous camera views [22,53]. Foreground blobs are shown in *solid green*. Candidate head-top points are depicted as *small orange circles*. 3D head location hypotheses are shown back-projected as *larger blue circles*. The current Viterbi path is depicted as a green line

To allow effective background subtraction, background images are used when testing this algorithm on the CHIL lecture corpus. These images are derived by splicing frames from the development set together, so as to remove the lecturer. This process is performed automatically, based on development CHIL data, and is possible for the "CHIL03" and "CHIL04" sets, since they contain development and evaluation data from the same lectures (see also Sect. 5.1). Furthermore, and similarly to all trackers used in this work, spatial constraints about the lecturer's 3D location are utilized to improve performance.

5 Experiments on the CHIL corpus

We now proceed to evaluate the performance of the proposed tracking scheme on the CHIL lecture corpus and compare it to

alternative approaches. Before reporting results, we briefly describe the CHIL lecture corpus, its annotations, and the adopted evaluation metrics.

5.1 The CHIL lecture corpus

Our experiments are conducted on the CHIL database. This consists of three subsets, with a fourth set becoming available in early 2007.
(i) *CHIL03*: This first dataset was collected in 2003 at the smart room of Universität Karlsruhe, in Germany (UKA), and contains seven lectures, each split into two development and two evaluation segments of approximately five minutes duration each. This set will be referred to as the "CHIL03" dataset, and it has been used in internal CHIL consortium evaluations during the summer of 2004.
(ii) *CHIL04*: The second phase of data collection took again place at the UKA smart room in late 2004. This effort resulted in five lectures, split in a similar fashion to the "CHIL03" set into a total of ten development and ten evaluation segments, each five minutes in duration. This will be referred to as the "CHIL04" set and has been employed in internal CHIL consortium evaluations in January 2005.
(iii) *CHIL05*: The most recent set is significantly more diverse, containing 18 development and 24 evaluation segments of lectures collected at two smart rooms, one located at UKA and the second at the Istituto Trentino di Cultura (ITC), in Italy (see also Figs. 1, 2). The development and evaluation sets correspond to disjoint lectures. This set will be referred to as "CHIL05", and it has been used in the first international evaluation campaign on the "Classification of Events, Activities and Relationships" (CLEAR) in March, 2006 [5]. It should also be mentioned that this collection effort includes three additional recording sites, partners of the CHIL consortium, including IBM Research. These data however belong to the so-called "interactive-seminar" (or meeting) scenario, where the focus is to determine the location of all meeting participants, typically being less than six in total. This part has been excluded from our experiments, since we concentrate on tracking the lecturer.

All video data in the three sets have been recorded using four synchronous corner cameras at 15 Hz. The frame resolution is 640 × 480 pixels for the UKA site and 800 × 600 pixels at ITC. In terms of data annotations, visible face locations have been manually labeled in all frame views for every 1.0s for the "CHIL05" data and 0.67s for the "CHIL03" and "CHIL04" sets. Furthermore, the bounding boxes of such faces have been labeled in the "CHIL04" and "CHIL05" sets, with additional facial feature points (nose bridge and eyes) annotated in the latter. In all cases, the corresponding 3D head centroid location is also given, as derived by triangulating the face labels across camera views. Therefore, evaluation of tracking algorithms is possible at the instants with available ground truths (at 0.67s or 1.0s intervals) using appropriate metrics, as discussed next.

5.2 Evaluation metrics

A number of metrics are used in our experiments to benchmark performance of 3D-head and 2D-face tracking algorithms. All are computed by comparing algorithmic outputs (estimated 3D head centroid locations or face bounding boxes) to their corresponding annotated ground truths on the evaluation data sets. In particular, the following are employed for benchmarking 3D head tracking in Sect. 5.3:

(i) *3D error*: This corresponds to the mean Euclidean 3D distance in millimeters between the estimated and the ground truth position of the head centroid in 3D coordinates. An additional 3D metric has been deemed of interest, namely the percentage of time instants, where the 3D error is smaller than 300 mm. This is denoted by "% 3D err < 300" in Table 1.
(ii) *2D error*: This is the mean Euclidean 2D distance in mm between the projection on the smart room floor of the estimated 3D head center and that of the corresponding ground truth projection. Furthermore, "% 2D err < 300" is the percentage of time instants, where the 2D error is smaller than 300 mm.

The above metrics have been employed in the first two years of CHIL internal evaluations (datasets "CHIL03" and "CHIL04"). They have been subsequently modified as part of the CLEAR 2006 evaluation campaign on the "CHIL05" dataset, in order to allow multi-person tracking. Details can be found in [56].

Concerning the 2D face detection task, a total of five metrics have been identified by the CHIL consortium for use

Table 1 Comparison of 3D head-tracking performance of various algorithms on the CHIL evaluation sets of 2003 and 2004

	DPAS	DPAS-f	BGS	MMS	DPAS-d
["CHIL03"]					
3D err (mm)	140.0	270.2	278.4	253.9	1649.4
2D err (mm)	123.6	217.3	204.7	228.3	1230.7
3D err < 300	92.9%	82.5%	81.2%	84.6%	13.2%
2D err < 300	93.3%	84.3%	84.1%	85.3%	14.6%
["CHIL04"]					
3D err (mm)	155.2	267.4	480.3	467.4	1852.4
2D err (mm)	141.8	208.9	436.9	441.1	1635.1
3D err < 300	95.4%	83.6%	47.7%	78.9%	10.9%
2D err < 300	95.6%	85.7%	57.1%	80.7%	12.6%

Clearly, the proposed system (DPAS) performs best

in the CLEAR 2006 evaluations [5]. Results based on the following three are reported in Sect. 5.4:

(i) Percentage of *correctly detected faces* ("Corr"), namely the percentage of detected faces with hypothesis–reference face bounding-box centroid distance more than half the size of the reference face.
(ii) Percentage of *wrong face detections* ("Err"), accounting for false positives (including detections with hypothesis–reference bounding-box centroid distance larger than half the reference face size).
(iii) Percentage of *missed face detections* ("Miss").

In these metrics, the reference face size is defined as the average of height and width of the annotated bounding box.

5.3 3D head tracking results

In the 3D head tracking experiments, we concentrate on the "CHIL03" and "CHIL04" subsets of the corpus. As discussed above, these contain non-overlapping development and evaluation subsets that correspond to the same lectures. This fact allows the training of relatively accurate face detectors, since they cover the same lecturer population (akin to a "multi-subject" training/testing scenario, as opposed to the more challenging "speaker-independent" case).

On these sets, we compare a total of five tracking algorithms:

(i) *DPAS*: This is the proposed face-detection based scheme that uses *dynamic programming* and *adaptive subspace* tracking.
(ii) *DPAS-f*: This is a variation of the proposed scheme, where no *forgetting* mechanism is introduced in the adaptive subspace tracking stage, thus the influence of past distant observations is retained until re-initialization is triggered.
(iii) *DPAS-d*: This is a trivial variation of the proposed scheme, where no *drift detection* is present. The algorithm gets initialized at the beginning of the multi-camera video sequence, and remains in the tracking stage, with no re-initialization until the sequence ends.
(iv) *MMS*: This corresponds to the algorithm described in Sect. 4.1. It constitutes a face-detection driven approach with *motion* based foreground segmentation and *mean shift* tracking.
(v) *BGS*: This is the system presented in Sect. 4.2 that uses *background subtraction* and dynamic programming.

All above systems are run to always return a 3D head centroid location. In case the algorithm fails to do so (for example, failing to initialize, as discussed in Sect. 3.2), the returned location defaults to the middle of the presenter's area or the previous estimate in time, if available (DPAS and MMS systems), or an interpolated location between existing estimates immediately before and after the particular instant (BGS system). Concerning face-detection based schemes, development set data are used to train the frontal and profile FloatBoost based face detectors, as discussed in Sect. 3.5. Furthermore, other system parameters, such as spatial constraints (all methods), DP costs (see for example Sect. 3.2), inter-ray distance thresholds (e.g., Sects. 3.2, 3.4), and tracking template sizes (Section 3.3, among others) are empirically determined on development data.

Results based on the 3D/2D error metrics discussed in Sect. 5.2 are depicted in Table 1. It is clear that the proposed system (DPAS) significantly outperforms all others. Interstingly, both systems described in Sect. 4 (MMS and BGS) achieve similar performance, but exhibit approximately twice (for the "CHIL03" set) or three times (for "CHIL04") the error of the proposed scheme. As expected, the variant of the proposed system, where no drift detection is present (DPAS-d), fails miserably. Finally, it is important to note that the introduction of the forgetting mechanism in adaptive subspace tracking plays a significant role in improving performance. This becomes clear from Table 1, since removing this component (DPAS-f system) almost doubles the tracking error (over DPAS). This is also illustrated in Fig. 8, where the evolution of 3D tracking error over time (quad-frame number) is depicted for one lecture segment.

Based on its superior performance on "CHIL03" and "CHIL04" data, the DPAS system was used for the CLEAR 2006 evaluation on the "CHIL05" set. A slight modification to the system was introduced, namely to return no 3D

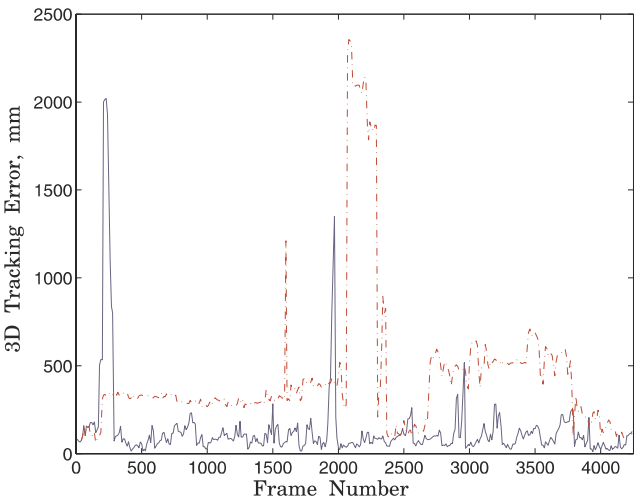

Fig. 8 Typical tracking behavior of the proposed system (DPAS: *solid line*), compared with its variant (DPAS-f: *dashed line*) with no forgetting mechanism, evaluated over a CHIL lecture segment

hypothesis when its initialization fails (see Sect. 3.2). This was deemed necessary due to the modified performance metrics in CLEAR 2006 that penalize guessing [56]. The resulting performance of the DPAS system on the "CHIL05" set was an average 2D error of 139.1 mm and 3D error of 145.5 mm. These numbers are very close to the ones achieved on "CHIL03" and "CHIL04" (see Table 1), demonstrating that the method generalizes well. A comparison of the DPAS with its DPAS-f variant on the same set shows that the latter exhibits significantly more tracking drifts, on the average every 193.9 quad-frames (instants), as opposed to 241.6 of the proposed DPAS tracker.

Additional comparisons of the proposed DPAS scheme with six alternative systems can be found in [5, pp. 29], as part of the CLEAR 2006 official evaluation. These systems have been briefly overviewed in Sect. 2.

5.4 2D face localization results

In the final set of experiments, we report the performance of the proposed 2D face localization subsystem, based on the DPAS head tracking system, as discussed in Sect. 3.1. The results are reported on the "CHIL05" set, used in the CLEAR evaluation campaign (see also [5, pp. 34]).

A summary of system performance based on the metrics of Sect. 5.2 is given in Table 2. The system achieved 54.5% correct detections, with 37.2% erroneous detections and 18.9% misses. This performance can be considered relatively good, if one takes into account the extremely challenging nature of the task and the rather strict evaluation metrics. In particular, by comparing the UKA development and evaluation set performance in Table 2, one can notice that the performance drops significantly, due to the different lecturer population sets (a purely "speaker independent" evaluation framework is considered). Furthermore, errors and misses are relatively balanced on the development set, but not so on the evaluation data.

Table 2 Performance of 2D face tracking on the "CHIL05" development (DEV) and evaluation (EVA) sets, depicted per collection site and cumulatively

"CHIL05" Data			Metrics (%)		
Set	Site	#Sem	Corr	Err	Miss
D	ITC	1	–	–	–
E	UKA	18	74.17	21.04	15.18
V	all	19	–	–	–
E	ITC	2	84.75	28.70	3.14
V	UKA	24	52.64	37.68	19.89
A	all	26	54.44	37.18	18.95

Number of seminar segments are also listed. All metrics are expressed in %

A final remark concerns the adopted strategy described in Sect. 3.1 for face detection. A number of approaches have been considered for producing 2D face results from the 3D head location estimate in an effort to reduce and balance the false positive ("Err") and negative ("Miss") error rates. Among them, an interesting modification of the proposed method is to always return the 2D tracking result on the two selected camera views where the subspace tracking takes place (Sect. 3.3), and only apply multi-pose face detection to the two non-tracked camera views around a region of interest based on the 3D head estimate. This is in contrast to first applying the multi-pose face detector on all four views, and only resorting to the tracking result of the selected camera views when the detector fails to return a face. The performance of the former approach was measured on seven UKA development set seminars at 77.26% Corr, 18.67% Err, and 9.37% Miss, compared to the superior 85.92% Corr, 9.95% Err, and 9.43% Miss of the adopted approach.

5.5 System run-time performance

There has been no particular effort to optimize the proposed system implementation. To reduce face detection overhead and allow speedier development, the whole system has been implemented in a cascade, where face detection is first applied at all instants and all camera views (as in Sect. 3.5), before feeding its output to the remaining system modules (described in Sects. 3.2–3.4). In practice, this is of course suboptimal, as the two 2D tracking processes (Sect. 3.3) can perform most of the required work in real time –20 f/s (frames per second) on a P4 2.8 GHz, 512 MByte desktop. In contrast, face detection over the entire frame in four camera views is significantly slower and runs only at about 2 f/s.

6 Summary and Discussion

In this paper, we have presented a vision system for joint 3D head and 2D face tracking for multi-camera smart room settings, where calibrated cameras with wide, overlapping fields of view synchronously record human interaction. In particular, the system has been developed for single-person tracking of the presenter in the CHIL lecture scenario. We described details of the system components, with important highlights being the use of AdaBoost-like multi-pose face detectors, employment of a spatio-temporal dynamic programming algorithm to initialize 3D location hypotheses, and the use of an adaptive subspace learning based 2D tracking scheme with a forgetting mechanism, as a means to reduce tracking drift and increase robustness. The proposed system deviates significantly from other literature work, by not relying on motion estimation, background subtraction, or human body appearance modeling.

We have extensively tested the system on three releases of the CHIL lecture corpus. The proposed system exhibited excellent results with 3D average tracking errors of 140, 155, and 146 mm on three test sets, and outperformed a number of competitive techniques considered in this paper, ranging from simple system variants to entirely different approaches. These experiments, as well as results of the CLEAR 2006 evaluation campaign, demonstrate that the proposed approach is well suited to the problem.

Nevertheless, the system has potential limitations: For example, it is clearly inappropriate for room/camera configurations that consistently result in capturing faces in a resolution too small to allow their detection. A second issue concerns extending the framework to multi-person tracking. Clearly, its 2D tracking and 3D drift detection modules are readily applicable to the multi-person task. However, robust redesign of the initialization module is more challenging. For this purpose, a dynamic programming framework that produces multiple tracks is envisaged, with the number of retained tracks optimized by ad-hoc or information-theoretic approaches.

In future work, we plan to continue research on the topic by working on the multi-person tracking problem. An additional area of interest concerns exploring appropriate multi-camera fusion schemes to allow the system tracking component to directly operate in the 3D space. A more efficient implementation in order to achieve faster run-time performance is also among our goals.

Acknowledgments This work was partially supported by the European Commission under integrated project CHIL, "Computers in the Human Interaction Loop", contract number 506909. The authors would like to acknowledge contributions by a number of IBM colleagues to the development of the background subtraction tracker used in this paper, through their work on the IBM Smart Surveillance Engine (SSE); in particular, Ruud Bolle, Lisa Brown, Jonathan Connell, Norman Haas, Arun Hampapur, Sharath Pankanti, and Ying-Li Tian. In addition, contributions by UIUC colleagues Jilin Tu and Ming Liu to the proposed system development are also acknowledged.

References

1. CHIL: Computers in the human interaction loop [Online]. Available: `http://chil.server.de`
2. Mostefa, D., Moreau, N., Choukri, K., Potamianos, G., Chu, S.M., Tyagi, A., Casas, J.R., Turmo, J., Christoforetti, L., Tobia, F., Pnevmatikakis, A., Mylonakis, V., Talantzis, F., Burger, S., Stiefelhagen, R., Bernardin, K., Rochet, C.: The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms. J. Lang. Resour. Eval. (submitted) (2007)
3. Stiefelhagen, R., Garofolo, J. (eds.) Multimodal Technologies for Perception of Humans: First International Evaluation Workshop on Classification of Events, Activities, and Relationships, CLEAR 2006. vol. 4122, Springer, LNCS (2007)
4. Fiscus, J.G., Ajot, J., Michel, M., Garofolo, J.S.: The rich transcription 2006 spring meeting recognition evaluation. In: Renals, S., Bangio, S., Fiscus, J.G. (eds.) Machine Learning for Multimodal Interaction, LNCS vol. 4299, pp. 309–322 (2006)
5. Stiefelhagen, R., Bernardin, K., Bowers, R., Garofolo, J., Mostefa, D., Soundararajan, P.: The CLEAR 2006 evaluation. In: Stiefelhagen, R., Garofolo, J. (eds.) Multimodal Technologies for Perception of Humans: First International Evaluation Workshop on Classification of Events, Activities, and Relationships, CLEAR 2006. vol. 4122, Springer, LNCS (2007), pp. 1–44 (2007)
6. Stergiou, A., Pnevmatikakis, A., Polymenakos, L.: A decision fusion system across time and classifiers for audio-visual person identification. In: Stiefelhagen, R., Garofolo, J. (eds.) Multimodal Technologies for Perception of Humans: First International Evaluation Workshop on Classification of Events, Activities, and Relationships, CLEAR 2006. vol. 4122, Springer, LNCS (2007), pp. 223–232 (2007)
7. Wölfel, M., Nickel, K., McDonough, J.: Microphone array driven speech recognition: influence of localization on the word error rate. In: Proceedings joint workshop on multimodal interaction and related machine learning algorithms (MLMI), LNCS vol. 3869, pp. 320–331 (2005)
8. Pinhanez, C., Bobick, A.: Intelligent studios: using computer vision to control TV cameras. In: Proceedings Workshop on Entertainment and AI/Alife, pp. 69–76 (1995)
9. Wallick, M.N., Rui, Y., He, L.: A portable solution for automatic lecture room camera management. In: Proceedings International Conference Multimedia Expo (ICME) (2004)
10. Hampapur, A., Pankanti, S., Senior, A.W., Tian, Y.-L., Brown, L., Bolle, R.: Face cataloger: multi-scale imaging for relating identity to location. In: Proceedings IEEE conference advanced video signal based surveillance, pp. 13–20 (2003)
11. Potamianos, G., Lucey, P.: Audio-visual ASR from multiple views inside smart rooms. In: Proceedings International Conference Multisensor Fusion and Integration for Intelligent Systems (MFI), pp. 35–40 (2006)
12. Bouguet, J.-Y.: Camera Calibration Toolbox [Online]. Available: http://www.vision.caltech.edu/bouguetj/calib_doc/
13. Pnevmatikakis, A., Polymenakos, L.: 2D person tracking using Kalman filtering and adaptive background learning in a feedback loop. In: Stiefelhagen, R., Garofolo, J. (eds.) Multimodal Technologies for Perception of Humans: First International Evaluation Workshop on Classification of Events, Activities, and Relationships, CLEAR 2006. vol. 4122, Springer, LNCS (2007), pp. 151–160 (2007)
14. Nechyba, M.C., Schneiderman, H.: PittPatt face detection and tracking for the CLEAR 2006 evaluation. In: Stiefelhagen, R., Garofolo, J. (eds.) Multimodal Technologies for Perception of Humans: First International Evaluation Workshop on Classification of Events, Activities, and Relationships, CLEAR 2006. vol. 4122, Springer, LNCS (2007), pp. 161–170 (2007)
15. Bernardin, K., Gehrig, T., Stiefelhagen, R.: Multi- and single view multiperson tracking for smart room environments. In: Stiefelhagen, R., Garofolo, J. (eds.) Multimodal Technologies for Perception of Humans: First International Evaluation Workshop on Classification of Events, Activities, and Relationships, CLEAR 2006. vol. 4122, Springer, LNCS (2007), pp. 81–92 (2007)
16. Nickel, K., Gehrig, T., Stiefelhagen, R., McDonough, J.: A joint particle filter for audio-visual speaker tracking. In: Proceedings International Conference Multimodal Interfaces (ICMI) (2005)
17. Abad, A., Canton-Ferrer, C., Segura, C., Landabaso, J.L., Macho, D., Casas, J.R., Hernando, J., Pardàs, M., Nadeu, C.: UPC audio, video and multimodal person tracking systems in the CLEAR evaluation campaign. In: Stiefelhagen, R., Garofolo, J. (eds.) Multimodal Technologies for Perception of Humans: First International Evaluation Workshop on Classification of Events, Activities, and Relationships, CLEAR 2006. vol. 4122, Springer, LNCS (2007), pp. 93–104 (2007)

18. Brunelli, R., Brutti, A., Chippendale, P., Lanz, O., Omologo, M., Svaizer, P., Tobia, F.: A generative approach to audio-visual person tracking. In: Stiefelhagen, R., Garofolo, J. (eds.) Multimodal Technologies for Perception of Humans: First International Evaluation Workshop on Classification of Events, Activities, and Relationships, CLEAR 2006. vol. 4122, Springer, LNCS (2007), pp. 55–68 (2007)
19. Wu, B., Singh, V.K., Nevatia, R., Chu, C.-W.: Speaker tracking in seminars by human body detection. In: Stiefelhagen, R., Garofolo, J. (eds.) Multimodal Technologies for Perception of Humans: First International Evaluation Workshop on Classification of Events, Activities, and Relationships, CLEAR 2006. vol. 4122, Springer, LNCS (2007), pp. 119–126 (2007)
20. Zhang, Z., Potamianos, G., Senior, A., Chu, S., Huang, T.: A joint system for person tracking and face detection. In: Proceedings International Workshop Human-Computer Interaction (ICCV 2005 Work. on HCI), pp. 47–59 (2005)
21. Lim, J., Ross, D., Lin, R.-S., Yang, M.-H.: Incremental learning for visual tracking. In: Proceedings NIPS (2004)
22. Hampapur, A., Brown, L., Connell, J., Ekin, A., Haas, N., Lu, M., Merkl, H., Pankanti, S., Senior, A., Shu, C.-F., Tian, Y.-L.: Smart Video Surveillance. IEEE Signal Process. Mag. **22**(2), 38–51 (2005)
23. Isard, M., MacCormick, J.: BraMBLe: A Bayesian multiple blob tracker. In: Proceedings International Conference Computer Vision, vol. 2, pp. 34–41 (2003)
24. Senior, A.: Real-time articulated human body tracking using silhouette information. In: Proceedings Workshop Visual Surveillance/PETS (2003)
25. Rowley, H.A., Baluja, S., Kanade, T.: Neural network-based face detection. IEEE Trans. Pattern Anal. Mach. Intell. **20**(1), 23–28 (1998)
26. Osuna, E., Freund, R., Girosi, F.: Training support vector machines: an application to face detection. In: Proceedings Conference Computer Vision Pattern Recog, pp. 130–136 (1997)
27. Roth, D., Yang, M.-H., Ahuja, N.: A SNoW-based face detector. In: Proceedings of NIPS (2000)
28. Viola, P., Jones, M.: Robust real time object detection. In: Proceedings IEEE ICCV Work. Statistical and Computational Theories of Vision (2001)
29. Graf, H.P., Cosatto, E., Potamianos, G.: Robust recognition of faces and facial features with a multi-modal system. In: Proceedings International Conference Systems Man Cybernetics pp. 2034–2039 (1997)
30. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. **23**(6), 681–685 (2001)
31. Pentland, A.P., Moghaddam, B., Starner, T.: View-based and modular eigenspaces for face recognition. In: Proceedings Conference Computer Vision Pattern Recogonition pp. 84–91 (1994)
32. Li, S.Z., Zhang, Z.: FloatBoost learning and statistical face detection. IEEE Trans. Pattern Anal. Mach. Intell. **26**(9), 1112–1123 (2004)
33. Isard, M., Blake, A.: Contour tracking by stochastic propagation of conditional density. In: Proceedings European Conference Computer Vision, pp. 343–356 (1996)
34. Comaniciu, D., Ramesh, V., Meer, P.: Real-time tracking of non-rigid objects using mean shift. In: Proceedings International Conference Computer Vision Pattern Recogonition vol. 2, pp. 142–149 (2000)
35. Tao, H., Sawhney, H.S., Kumar, R.: Dynamic layer representation with applications to tracking. In: Proceedings International Conference Computer Vision Pattern Recogonition vol. 2, pp. 134–141 (2000)
36. Black, M.J., Jepson, A.: Eigentracking: robust matching and tracking of articulated objects using a view-based representation. Int. J. Comput. Vis. **26**(1), 63–84 (1998)
37. Jepson, A.D., Fleet, D.J., El-Maraghi, T.F.: Robust online appearance models for visual tracking. IEEE Trans. Pattern Anal. Mach. Intell. **25**(10), 1296–1311 (2003)
38. Collins, R.T., Liu, Y., Leordeanu, M.: Online selection of discriminative tracking features. IEEE Trans. Pattern Anal. Mach. Intell. **27**(10), 1631–1643 (2005)
39. Han, B., Davis, L.: On-line density-based appearance modeling for object tracking. In: Proceedings International Conference Computer Vision (2005)
40. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision 2nd edn. Cambridge University Press, ISBN: 0521540518 (2004)
41. Lanz, O.: Approximate Bayesian multibody tracking. IEEE Trans. Pattern Anal. Mach. Intell. **28**(9), 1436–1449 (2006)
42. Zotkin, D.N., Duraiswami, R., Davis, L.S.: Joint audio-visual tracking using particle filters. EURASIP J. Appl. Signal Process. **2002**(11), 1154–1164 (2002)
43. Mittal, A., Davis, L.: M2Tracker: a multi-view approach to segmenting and tracking people in a cluttered scene using region-based stereo. In: Proceedings European Conference Comp. Vision, pp. 18–36 (2002)
44. Kalman, R.E.: A new approach to linear filtering and prediction problems. Trans. ASME J. Basic Engin. (Ser. D) **82**, 35–45 (1960)
45. Arulampalam, M.S., Maskell, S., Gordon, N., Clapp, T.: A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. IEEE Trans. Signal Process. **50**(2), 174–188 (2002)
46. Stauffer, C., Grimson, W.E.L.: Learning patterns of activity using real-time tracking. IEEE Trans. Pattern Anal. Mach. Intell. **22**(8), 747–757 (2000)
47. Tyagi, A., Potamianos, G., Davis, J.W., Chu, S.M.: Fusion of multiple camera views for kernel-based 3D tracking. In: Proceedings IEEE Workshop Motion and Video Computing (2007)
48. Ho, J., Lee, K.-C., Yang, M.-H., Kriegman, D.: Visual tracking using learned linear subspaces. In: Proceedings International Conference Computer Vision Pattern Recogonition. vol. 1, pp. 782–789 (2004)
49. Hall, P., Marshall, D., Martin, R.: Merging and splitting eigenspace models. IEEE Trans. Pattern Anal. Mach. Intell. **22**(9), 1042–1049 (2000)
50. Freund, Y., Schapire, R.: A decision-theoretic generalization of on-line learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997)
51. Tieu, K., Viola, P.: Boosting image retrieval. In: Proceedings Conference Computer Vision Pattern Recogonition vol. 1, pp. 228–235 (2000)
52. Pudil, P., Novovicova, J., Kittler, J.: Floating search methods in feature selection. Pattern Recog. Lett. **15**, 1119–1125 (1994)
53. Senior, A.W., Potamianos, G., Chu, S., Zhang, Z., Hampapur, A.: A comparison of multicamera person-tracking algorithms. In: Proceedings IEEE International Workshop Visual Surveillance (VS/ECCV) (2006)
54. Bobick, A., Davis, J.: The representation and recognition of action using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. **23**(3), 257–267 (2001)
55. Senior, A.: Tracking with probabilistic appearance models. In: Proceedings International Workshop on Performance Evaluation of Tracking and Surveillance Systems (2002)
56. Bernardin, K., Elbs, A., Stiefelhagen, R.: Multiple object tracking performance metrics and evaluation in a smart room environment. In: Proceedings IEEE International Workshop Visual Surveillance (VS/ECCV) (2006)