Face Tracking and Recognition with Visual Constraints in Real-World Videos

Minyoung Kim[1], Sanjiv Kumar[2], Vladimir Pavlovic[1], and Henry Rowley[3]
[1]Dept. of Computer Science, Rutgers University
Piscataway, NJ 08854
{mikim,vladimir}@cs.rutgers.edu
[2]Google Research
New York, NY 10011
sanjivk@google.com
[3]Google Research
Mountain View, CA 94043
har@google.com

Abstract

We address the problem of tracking and recognizing faces in real-world, noisy videos. We track faces using a tracker that adaptively builds a target model reflecting changes in appearance, typical of a video setting. However, adaptive appearance trackers often suffer from drift, a gradual adaptation of the tracker to non-targets. To alleviate this problem, our tracker introduces visual constraints using a combination of generative and discriminative models in a particle filtering framework. The generative term conforms the particles to the space of generic face poses while the discriminative one ensures rejection of poorly aligned targets. This leads to a tracker that significantly improves robustness against abrupt appearance changes and occlusions, critical for the subsequent recognition phase. Identity of the tracked subject is established by fusing pose-discriminant and person-discriminant features over the duration of a video sequence. This leads to a robust video-based face recognizer with state-of-the-art recognition performance. We test the quality of tracking and face recognition on real-world noisy videos from YouTube as well as the standard Honda/UCSD database. Our approach produces successful face tracking results on over 80% of all videos without video or person-specific parameter tuning. The good tracking performance induces similarly high recognition rates: 100% on Honda/UCSD and over 70% on the YouTube set containing 35 celebrities in 1500 sequences.

1. Introduction

Despite recent progress, accurate face recognition remains a challenging task in dynamic environments, such as video, where noise conditions, illumination, and the subject's location and pose can vary significantly from frame to frame. At the same time, video-based recognition provides a setting where weak evidence in individual frames can be integrated over long runs of video, potentially leading to more accurate recognition in spite of the added difficulties. In this paper, we present a new method for face tracking and recognition in video that successfully circumvents the difficulties and leverages the benefits of the video setting and is capable of dealing with unconstrained real-world videos.

Effectively solving the video-based face recognition problem depends on two tasks: accurate face tracking and interpretation/classification of the tracked data. Face tracking is a critical prior step that localizes the region of the face in video frames, from which a relevant feature set can be extracted and subsequently served as input to the face recognizer. As such, the accuracy of tracking directly impacts the ability to recognize subjects in video.

Visual tracking of objects of interest, such as faces, has received significant attention in the vision community. Accurate tracking is made difficult by the changing appearance of targets due to their nonrigid structure, 3D motion, interaction with other objects (*e.g.*, occlusions) and changes in the environment, such as illumination. Recent tracking methods, such as the Incremental Visual Tracker (IVT) [18], attempt to solve these problems using adaptive target appearance models. They represent the target in a low-dimensional subspace which is updated adaptively using the images tracked in the previous frames. Compared to the approaches equipped with a fixed target model such as eigentracking of [4], IVT is more robust to changes in appearance (*e.g.*, pose, illumination). However, the main drawback of the adaptive approaches is their susceptibility to drift: they can gradually adapt to non-targets as the target model is built solely from the previous tracked images accepted by the tracker. Methods such as IVT typically lack mechanisms for detecting or correcting drift as they have no global constraints on the overall appearance of the target object. For faces, such constraints could be learned from a set of generic (non-person specific) well-cropped and well-aligned face images that span possible variations in pose, illumination, and expressions. These can be seen as *visual constraints* that the target appearance should meet.

To achieve this, our tracker introduces two new constraint terms to the adaptive target model in a particle filtering framework: (1) a generative model based set of facial pose subspaces or manifolds, each of which represents

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

a particular out-of-plane pose, and (2) and a discriminative model (SVM) based goodness-of-crop discriminator whose confidence score indicates how well the cropped face is aligned. These constraint terms are linked with the adaptive term of IVT, leading to a new constrained 'likelihood' potential in the state-space model. We demonstrate that this new tracker significantly improves robustness against occlusion and abrupt appearance change.

Recognition of people's faces in video can be done in a static, frame-by-frame fashion [1, 16]. However, a dynamic setting provides additional constraints that can increase the accuracy of recognition [5, 13, 17, 19]. Heuristic temporal voting schemes such as [3, 15] aggregate data from key frames containing well-illuminated frontal poses. This makes the performance of these approaches sensitive to the quality (or even existence) of such key frames. An alternative approach is to rely on the full sequence of well-tracked frames to yield a final recognition decision [17]. The accuracy of recognition in this setting is critically related to the ability to effectively discriminate between multiple face poses. Features such as PCA used in [17] may not be sufficiently discriminative, especially when the number of subjects is large. Features based on discriminative spaces such as those produced by LDA (Linear Discriminant Analysis), coupled with properly modeled pose dynamics, can lead to significant improvements in the recognition accuracy. In this work we show that both of these aspects can be modeled within an HMM-based recognition framework by explicitly guiding the hidden states in HMMs to be facial poses. When coupled with a proper, well-constrained tracking solution, this leads to state-of-the-art recognition performance.

As the main contribution of this work, we show that state-of-the-art adaptive trackers (*e.g.*, IVT) can be made significantly more robust to occlusion and illumination variation by adding *non-adaptive*, *non-person specific* constraints on face pose and localization. Through extensive experiments on the full Honda/UCSD subjects and a new YouTube 35-subject, 1500-sequence set, we demonstrate that these constraints are not only significant for tracking, but also critical for subsequent recognition. We also show that, in the absence of labeled tracking data, performance of trackers can be quantitatively measured via face recognition.

The rest of the paper is organized as follows. A brief review of current video-based recognition approaches is presented in Section 2. We then describe our new constrained adaptive face tracker and show its basic advantages. Section 4 introduces the video-based recognition framework that relies on the tracker in Section 3. We finally demonstrate, in an extensive set of experiments, the performance of the coupled tracking-recognition framework on Honda/UCSD data as well as a large video database of YouTube video clips of 35 celebrities. Our approach produces successful face tracking results on a large fraction of the videos without instance-specific parameter tuning, while achieving high recognition rates.

2. Prior Work

In this section we review recent approaches to tracking and video-based face recognition. While many different approaches have been proposed in the past we briefly focus on those most related to our approach.

Robustness of tracking and adaptation to changing target appearance and scene conditions are critical properties a tracker should satisfy. Numerous approaches to target modeling have attempted to tackle these issues using view-based appearance models [4], contour models [9], 3D models [11], mixture models [10], and kernel representations [6, 8], among others. Direct use of object detectors in discriminative tracking has been proposed more recently, *c.f.*, [2, 14]. For instance, [14] tackled the challenging case of the low frame rate videos by integrating tracking and detection in a cascade fashion, where the lifespan and feature sets of observation models change during the tracking process to increase efficiency and robustness of the tracker. An adaptive graph-based discriminative tracker [20] combines foreground templates with updating background model. However, learning a model for discrimination from the full background is usually difficult. Furthermore, the problem of small drifts from the face, *i.e.*, the "goodness of the crop", critical for the recognition stage, is typically not addressed in these approaches. In contrast, our proposed tracker is specifically designed to allow adaptation of the model while, at the same time, adhering to the global object class (face) using visual constraints. Moreover, the tracker is required to be subject-agnostic and generic in the setting of specific tracking parameters, allowing it to be easily applied to a large body of diverse videos.

Coupling of face tracking and recognition has attracted interest in the vision community. For instance, a state-space model was proposed in [5] for classifying videos where state dynamics separates different subjects. [13] assumed that the appearance of faces lies on a probabilistic manifold specific to a subject's identity, approximated by a set of pose-specific linear subspaces. Subject identification in a video sequence is accomplished by finding the closest manifold (identity) where distance is computed using a temporal fusion in a Bayesian fashion. However, this approach suffers from the need for off-line trained subject-specific trackers, which increases the number of model parameters that need to be set preventing scalability of such an approach. This approach was extended to simultaneously deal with tracking and recognition [12] using an initial generic appearance manifold that is adapted, in the course of tracking, to a person-specific manifold. The on-line adaptation

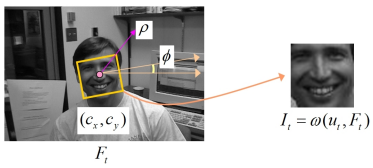


Figure 1. The warping function from the tracking state $u_t = [c_x, c_y, \rho, \phi]^T$ (solid box) to a cropped image I_t.

process, however, relies on training using synthesized face images specific to the currently predicted identity which may limit this approach to situations where such models are available. Very recently, [19] considered the face recognition problem in real-world videos, containing uncontrolled variations in facial appearance. They accomplish this by assigning confidence scores from local classifiers to the face images in each frame, and then obtaining the sequence-level prediction using heuristic weighting of frame-based confidences. Despite promising results, the need for significant parameter tuning and heuristic integration schemes may limit the generalization of this approach.

3. Face Tracking

In a probabilistic framework, tracking can be seen as (online) temporal filtering that estimates:

$$P(u_t|F_{0\ldots t}), \quad \text{for } t = 1, 2, \ldots, \tag{1}$$

where F_t is the input image frame and u_t is the tracking state at time t. The initial state u_0 is assumed to be known. In this paper, we use similarity transformation parameters $u_t = [c_x, c_y, \rho, \phi]^T$, where the first two elements are the center position of the square tracking box, ρ is the scale w.r.t. the standard image size (48×48), and ϕ is the in-plane rotation angle from the horizontal axis (Fig. 1). The tracker is required to localize the face in space (c_x,c_y) as well as in size and orientation. Accurate estimation of all four parameters is crucial for subsequent use of the detected image in the face recognition phase which is typically sensitive to alignment. Given F_t, the tracking state u_t determines the cropped face image I_t by the warping function $I_t = \omega(u_t, F_t)$, as illustrated in Fig. 1.

The filtering of (1) typically assumes a 1st-order state-space model depicted in Fig. 2. Smoothness of tracking is enforced by the temporal dynamics. While various dynamic models could be used, motion without gross changes typically justifies the use of a simple Gaussian smoothness, namely,

$$P(u_t|u_{t-1}) = \mathcal{N}(u_t; u_{t-1}, \Sigma), \tag{2}$$

with a proper choice of Σ. The emission model effectively takes the warped cropped image $I_t = \omega(u_t, F_t)$ as an observation feature and evaluates its score w.r.t. the underlying

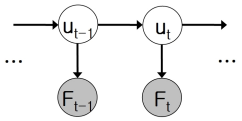


Figure 2. First order state-space model for tracking.

target appearance model. We define the emission probability as a Gibbs energy with scale σ

$$P(F_t|u_t) \propto e^{-E(\omega(u_t,F_t);\theta)/\sigma^2}, \tag{3}$$

where $E(I_t;\theta)$ is the energy function w.r.t tracker's target model θ, which attains lower values when I_t is more compatible with θ.

Under this model, (1) is obtained using the standard Bayesian recursion:

$$P(u_t|F_{0\ldots t}) \propto \int P(u_t|u_{t-1}) \cdot P(u_{t-1}|F_{0\ldots t-1}) du_{t-1} \times P(F_t|u_t). \tag{4}$$

As the integration in (4) is intractable analytically due to the non-Gaussian form of $P(F_t|u_t)$ (in terms of u_t), we resort to sampling-based particle filtering. A set of weighted particles $\{(w^i, u^i)\}_{i=1}^n$ is maintained to approximate the conditional density $P(u_{t-1}|F_{0\ldots t-1})$ at time $t-1$. These particles undergo dynamics $P(u_t|u_{t-1})$ followed by *re-weighting* according to $P(F_t|u_t)$ at time t.

The energy $E(I_t)$ plays a crucial role as it estimates the confidence of a candidate particle in terms of its quality. The simplest first-frame (or a two-frame) tracker has a fixed target model as the initial track I_0 (or the previous track I_{t-1}), which defines the energy as a distance to this template, namely, $E(I_t) = d(I_t, I_0)$ (or $E(I_t) = d(I_t, I_{t-1})$), where $d(\cdot,\cdot)$ is a distance measure in the image space. These simplistic energy functions typically make the tracker either too inflexible or too susceptible to appearance variations due to changes in 3D face orientation, scene lighting, occlusions, etc.

To cope with this, IVT [18] employs a more sophisticated target model that reflects changes in appearance. It builds a linear (PCA) subspace representation $M(I_{0\ldots t-1}) = (\mu, B)$ for the target, where the subspace mean μ and the basis B are updated by an incremental SVD on the previous tracks $I_0, \ldots, I_{t-1}$. In particular, $E(I_t)$ is defined as the reconstruction error of I_t w.r.t. $M(I_{0\ldots t-1})$

$$E(I_t; (\mu, B)) = ||(I_t - \mu) - BB^\top (I_t - \mu)||^2, \tag{5}$$

where B contains subspace bases as its columns.

3.1. Adaptive tracking with visual constraints

Although the adaptive property of IVT provides robustness to smooth changes in appearance, it may cause the

tracker to drift, i.e, move away from the desired target by gradually adapting to non-targets. One reason for this is that IVT lacks strong mechanisms for detection or correction of drifting.

One way to reduce the drift is to introduce additional constraints on the appearance of the tracked object. A model accounting for such constraints can be built from off-line data with enough variation in appearance. Since we restrict ourselves to the class of human *faces*, one can define a reasonable set of visual constraints that serves as an indicator for detection or correction of drifting. In this work we propose two such constraints; one for facial pose and the other for the alignment of the cropped faces.

3.1.1 Pose constraints

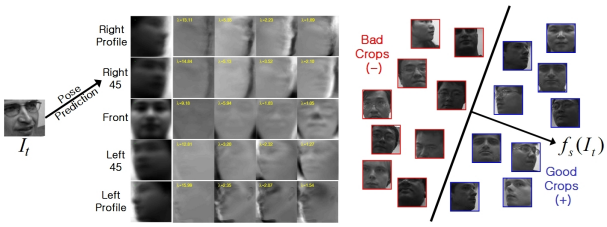
To constrain the appearance across different (out-of-plane) poses, we construct a set of pose subspaces. We consider a set of linear subspaces encapsulated in the model $M_p = \{(\mu_i, B_i)\}_{i=pose}$. Fig. 3(a) illustrates one such model. We then define the energy related to this pose constraint as a minimum distance among the pose subspaces, namely, $d(I_t, M_p) = \min_i d(I_t, (\mu_i, B_i))$. We use a reconstruction error similar to that of [18] as a distance measure. Intuitively, this term prevents the target from drifting away from predefined pose prototypes.

The pose subspace model can be estimated from a dataset of differently oriented face images. We used the face data from the Honda/UCSD video database in [12]. We detect faces and manually align them, obtaining about $8,000$ face images of different poses with all 14 different people and varying illumination conditions. We roughly categorize the poses into 5 clusters (*frontal*, *left/right* 45*-deg*, *left/right profile*)[1], and data from each pose cluster is used to train a PCA subspace, forming a set of pose subspaces.

3.1.2 Alignment constraints

The alignment constraint determines whether or not the candidate image contains a well-aligned and cropped face. Fig. 3(b) depicts some examples of correctly and incorrectly cropped faces. Determining how well an image is cropped can be accomplished using a confidence score of a classifier that discriminates well-cropped face images from the drifted images or, possibly, non-faces. In this case the face data (for all poses and subjects) used in pose subspace learning become positive examples for learning a classifier, such as an SVM. The ill-cropped negative images were obtained by shifting, rotating, and scaling the good examples randomly by a significant amount. We use $f_s(I_t)$ to denote the confidence of an SVM classifier learned in this manner.

[1]We ignored the poses in vertical direction (up/down) for simplicity.



(a) Pose subspaces (b) SVM crop classifier

Figure 3. Two visual constraints. (a) Each row represents the PCA subspace learned for each pose. The distances from I_t to the subspaces are computed, and the minimum is selected as a predicted pose for I_t. (b) SVM classifier that discriminates well-cropped face images ($+$, in blue) against drifted or shifted images ($-$, in red).

3.1.3 Visually constrained adaptive model

The two constraint terms are combined with an IVT-like adaptive term $M_a(I_{0\ldots t-1})$, weighted by contribution factors $\lambda_a, \lambda_p, \lambda_s > 0$, into the final energy function

$$E(I_t) = \lambda_a d(I_t, M_a(I_{0\ldots t-1})) + \lambda_p d(I_t, M_p) - \lambda_s f_s(I_t). \quad (6)$$

The intuition behind this function is appealing: the tracker assigns higher confidences to the particles that not only match the adaptive model, but also conform to the pre-specified generic facial pose and crop alignment constraints.

To illustrate the impact of the proposed visual constraint terms, we compared our tracker given in (6) with IVT, which is a special case of our model obtained by setting $\lambda_p = \lambda_s = 0$. Fig. 4 depicts one case where the lack of additional constraints induces a failure in the IVT. The image sequence exhibits occlusion in frames $t = 104 \sim 106$. Our tracker finds the correct target guided by the two constraint terms, while the IVT drifts from the target as it adapts to non-targets acquired during $t = 104 \sim 106$. This example signifies the importance of visual constraints for robust adaptive object tracking[2].

4. Video-based Face Recognition

Tracking provides well cropped/aligned face images $I_{1,\ldots,T}$ for face recognition. Fig. 5 shows examples of images obtained in the tracking phase that are used in this process. The task of the face recognition phase is to label an arbitrary video sequence with the identity of the person in the video clip. We assume a single subject in each clip. However, in the case of multiple people, recognition is not affected if we can provide multiple face tracks.

4.1. Face recognition using HMM

The video-based recognition task can be cast as the general problem of sequence classification. In this setting, we

[2]More examples showing advantages of adding each constrain term can be found at http://seqam.rutgers.edu/projects/motion/face/face.html.

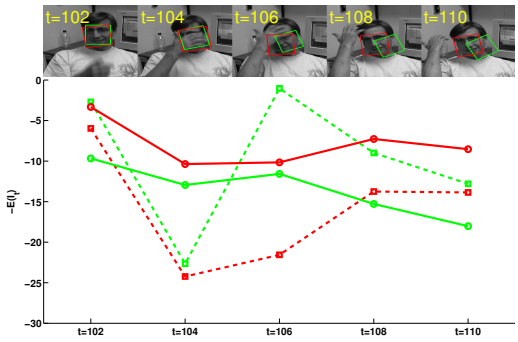


Figure 4. Example of tracking by proposed tracker and IVT: (*Top*) Face undergoing the occlusion at $t = 104 \sim 106$, IVT (green/light box) is adapted to the non-target images, while the proposed tracker (red/dark box) survives due to the two constraint terms (pose + svm). (*Bottom*) Tracked data compatibility (data log likelihood) of the two trackers. Lines in red (green) are the values of $-E(I_t)$ evaluated on the red (green) boxes by the proposed tracker (solid) and IVT (dashed). During the occlusion, IVT strongly adapted to the wrong target (*e.g.*, $t = 106$), leading to a highly peaked data score. Consequently, at $t = 108$, the green particle is incorrectly chosen as the best estimate. Visual constraints restrict the adaptation to the occluding non-target, producing more accurate hypotheses in the subsequent frames.

Figure 5. Example face sequences used for recognition.

use HMM as the modeling paradigm. Our face recognition model is shown in Fig. 6. The subject (class) variable $y \in \{1, \ldots, M\}$ is one of M subjects. The observation features of the model, denoted by x_t, are extracted from the image I_t by a feature extractor. Further, $s = s_1, \ldots, s_T$ denotes the hidden state sequence. Existence of good features and an appropriate choice of the hidden state are critical for sound recognition performance.

Facial pose presents an appealing choice for the hidden state. Unlike arbitrary PCA-based subspaces, the pose space may allow the use of well-defined discriminative pose features in the face recognition HMM. We pursue this approach in our work. In particular, s_t represents a particular pose among J possible poses, $s_t \in \{1, \ldots, J\}$, and x_t denotes a pose-discriminant feature vector described below. The appearance sequence of length T is then modeled by the generative model (HMM) for each subject y:

$$P_y(s, x) = P(s_1) \cdot \prod_{t=2}^{T} P(s_t|s_{t-1}) \cdot \prod_{t=1}^{T} P_y(x_t|s_t). \quad (7)$$

The pose discriminating features are obtained using the LDA (Linear Discriminant Analysis). Each image I_t is

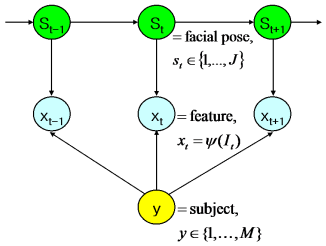


Figure 6. A graphical model for face recognition.

projected onto a discriminant subspace ψ trained by LDA on pose labeled data to yield $x_t = \psi(I_t)$. We used generic (not person-specific) face images from a subset[3] of Honda/UCSD face video database, and hand-labeled them in 7 different poses (up, down, L/R profiles, L/R 45-deg and front). This LDA training results in 6-dim representation of the face images.

The subject-specific observation density, needed by each HMM, is modeled as a Gaussian distribution, namely, $P_y(x_t|s_t = j) = \mathcal{N}(x_t; m_j^y, V_j^y)$, where m_j^y and V_j^y are the mean and the covariance for pose j of subject y, respectively. We let the pose dynamics be shared across all y's. This is a reasonable assumption, implying that the way poses change is generic and independent of any one particular person. We also verified experimentally on the YouTube dataset (Sec. 5.3) that pose dynamics were very similar. An added benefit of this assumption is the overall simplification of the model and reduction of the parameter set that needs to be estimated.

The model is trained by the EM algorithm with the subject labeled face sequence data. Note that we do not need the pose label at each frame as this can be inferred in the E-step of EM. At test time, for a new sequence $x_{1,\ldots,T}$, the subject estimation can be done in two different forms: the overall class prediction (*i.e.*, smoothing) and the on-line class estimation (*i.e.*, filtering). The class smoothing is:

$$y^* = arg\max_y P(y|x_{1,\ldots,T}) = arg\max_y P(y)P_y(x_{1,\ldots,T}), \quad (8)$$

and the on-line class filtering can be done recursively:

$$P(y|x_{1,\ldots,t+1}) \propto P(y|x_{1,\ldots,t}) \cdot \sum_{s_t, s_{t+1}} P_y(x_{t+1}|s_{t+1})P(s_{t+1}|s_t)P_y(s_t|x_{1,\ldots,t}), \quad (9)$$

where the last quantity $P_y(s_t|x_{1,\ldots,t})$ is the well-known forward state estimation for y's HMM. One may also be interested in the on-line pose estimation, *i.e.*, $P(s_t|x_{1,\ldots,t})$, which is similarly derived as:

$$P(s_t|x_{1,\ldots,t}) = \sum_y P_y(s_t|x_{1,\ldots,t})P(y|x_{1,\ldots,t}). \quad (10)$$

[3]The subjects in this off-line training set do not appear in any test data we used in the paper.

To demonstrate the benefit of LDA features, we compared our LDA based model to the one that utilizes PCA-based features of varying dimension, as in [17]. The recognition rates for the Honda/UCSD face videos are shown in Table 1. The result suggests that increasing the dimensionality of PCA improves the recognizer's performance. However, PCA features remain inferior to the pose discriminating LDA feature of substantially lower dimensionality.

Table 1. The recognition accuracies of the proposed model (Fig. 6) on the Honda/UCSD dataset. Either the pose discriminating LDA features (6 dimensions) or the PCA features (with varying dimensions) are used as observation features. Increasing the PCA dimension over 50 results in overfitting.

Feature	Accuracy	Feature	Accuracy
LDA (6-dim)	97.62 %	PCA (20-dim)	88.10 %
PCA (5-dim)	85.71 %	PCA (30-dim)	88.10 %
PCA (10-dim)	88.10 %	PCA (50-dim)	92.86 %

Another interesting insight into the model performance can be obtained by considering how its predictions of pose and subject ID change in the course of time. We illustrate this in Fig. 7. It shows that improved prediction performance results from acquisition of additional, accurately tracked subject face frames despite changes in pose, a condition that challenges many traditional face recognition systems.

4.2. Incorporating landmark-based features

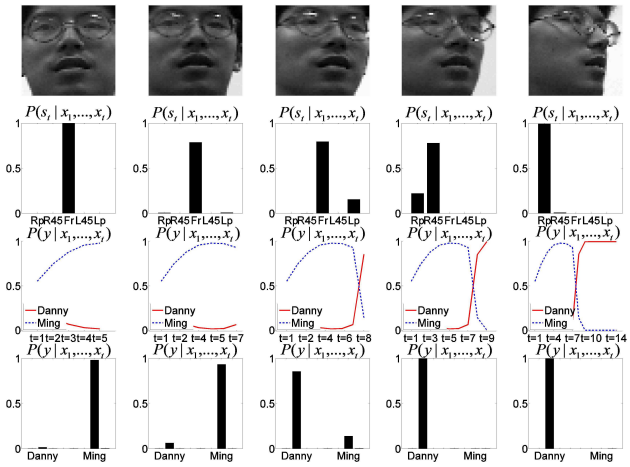
Sec. 4.1 has demonstrated the benefits of the LDA features on discriminating pose. For recognition, in addition to these, we use *LMT* (*i.e.* LandMark Template) features. The LMT features consist of multi-scale Gabor features (at 6 scales and 12 orientations) applied to 13 automatically located landmarks within the face bounding box. Since the LMT features are high (~ 1000) dimensional, we used PCA to extract only 10 major factors. We concatenate the LMT features with the pose-discriminating LDA features to form an observation feature vector for our recognizer.

5. Experiments

We first evaluate the proposed approach on benchmark datasets specifically designed for evaluating the performance of tracking (Sec. 5.1) and recognition (Sec. 5.2). In Sec. 5.3, we conduct both tracking and recognition experiments on a new set of challenging real-world YouTube videos.

5.1. Tracking on standard video datasets

Fig. 8 shows the tracking results of the IVT and our approach on example videos. The *trellis70* dataset from [18] exhibits severe illumination conditions with partial shading.



(a) t=5 (b) t=7 (c) t=8 (d) t=9 (e) t=14

Figure 7. Example of face recognition on Honda/UCSD dataset: The top row shows an example face sequence, the second row gives the pose prediction, $P(s_t|x_{1,...,t})$, and the bottom two rows depict the subject prediction, $P(y|x_{1...t})$, in historical and histogram views. The pose is predicted correctly changing from *frontal* to *R-profile*. The true class is *Danny*. It is initially incorrectly identified as Ming (blue/dashed curve in the third row). In subsequent frames the red/solid curve overtakes Ming, resulting in correct final prediction.

The other videos from IIT-NRC [7] contain relatively low quality image frames with abrupt changes in pose and size.

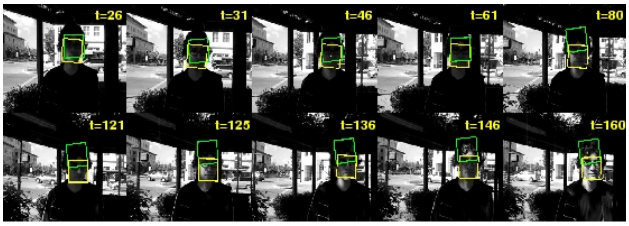
5.2. Recognition on Honda/UCSD dataset

Videos in this dataset include large variations in out-of-plane (left/right and up/down) head movement as well as in facial expression. The set contains several dozen subjects, each one appearing in at least two sessions. After applying our face tracker from Sec. 3.1, the sequences of cropped face images were used as input observations to the face recognizer.

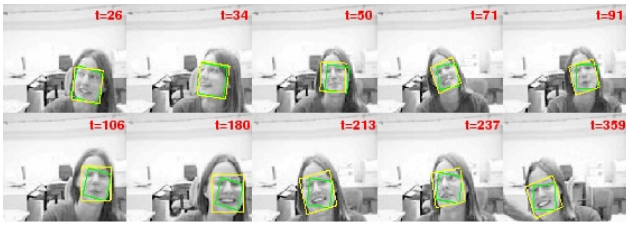
We followed the setting similar to that of [12, 13]. Table 2 shows recognition rates for several competing methods: Our model with LDA+LMT, LDA-only, and PCA-only features, the manifold-based approaches of [12, 13], and three standard frame-based methods. Our proposed approach with LDA features outperforms other state-of-the-art methods, while incorporating discriminative LMT features further improves the performance.

5.3. YouTube celebrity recognition

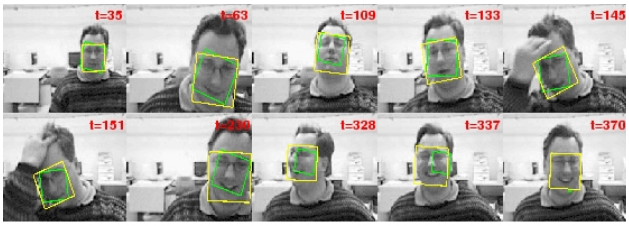
The proposed face tracking and recognition algorithms were also tested on a large set of noisy real-world videos. We collected video clips of 35 celebrities, mostly actors/actresses and politicians, from YouTube. Most of the videos are low resolution and recorded at high compression rates. This leads to noisy, low-quality image frames. As the video clips usually contain frames with no celebrity of in-



(a) Trellis 70



(b) NRC 08-2



(c) NRC 10-2

Figure 8. Tracking results on standard video datasets: The yellow is the proposed tracker and the green is the IVT. In (b) and (c), the IVT appears to track the position well, but the errors in size and rotation prevent the tracked frames from being used in recognition.

Table 2. Recognition accuracies on the Honda/UCSD dataset.

Method	Accuracy	Method	Accuracy
LDA + LMT	100.00 %	LDA Only	97.62 %
PCA (50-dim)	92.86 %	On-line [12]	95.60 %
Off-line [13]	97.20 %	Eigen-Faces	69.30 %
Fisher-Faces	74.50 %	Nearest Neighbor	81.60 %

terest, we manually segmented the clips into homogeneous sequences where the celebrity of interest does appear. The segmented dataset consists of about 1500 video clips, each one containing hundreds of frames. The frame sizes range from (180×240) to (240×320). This database[4] is challenging for face trackers and recognizers as the videos exhibit large variations in face pose, illumination, expression, and other conditions.

The proposed tracker was applied to the videos after manual marking of the initial state u_0 (or I_0). Alternatively, one could use a face detector to initialize the tracking. Unlike many trackers whose parameters are tuned to each individual sequence prior to tracking, our tracker's parameters

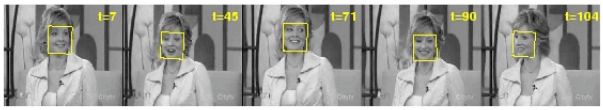
[4] Available on http://seqam.rutgers.edu/projects/motion/face/face.html.

(a) Tony Blair 02-001

(b) Al Pacino 02-02

(c) Angelina Jolie 01-005

(d) Julia Roberts 02-004



(e) Jodie Foster 03-006

(f) Steven Spielberg 03-009

(g) Bill Gates 03-005

Figure 9. Tracking results on YouTube datasets.

are *identical* for all 1500 sequences. Our tracker successfully tracked 80% of the video clips. Fig. 9 shows examples of the well-tracked videos. This performance level is remarkably high given the variability in pose, expression, size, and dynamics in this dataset. Moreover, the tracker implemented in Matlab was able to process $3 \sim 4$ frames per second with hundreds of particles.

For recognition, we randomly partitioned the well-tracked videos into train/test sets. As a baseline performance measure, we employed standard approaches based on key-frame selection. Such methods rely on the confidence scores assigned to each of the frames to select the most discriminative ones. For this purpose, we use the LMT features, where the confidence score for each frame is based on how well the landmark points match a pre-defined template. Two standard approaches are used in our experiment: *key-frame representative* - after selecting the frame with the highest confidence, we predict using the nearest neighbor in LMT feature space; and *key-frame voting* - a majority voting scheme is applied to the frames that have scores above a predefined threshold.

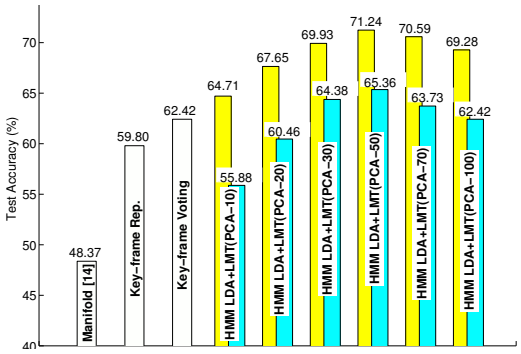


Figure 10. Celebrity recognition results on YouTube videos. The HMM-based face recognizer is tested on the tracking results of our face tracker (yellow/light) and the IVT (cyan/dark).

As shown in Fig. 10, the key-frame representative recorded about 60% accuracy (note that a random guess has $1/35 \sim 3\%$ accuracy) while voting improved the accuracy slightly. Note that key-frame based methods do significantly better than the manifold-based approach of [13], which had an accuracy of 48.37%. Incorporating the pose dynamics through an HMM, and a combination of LDA and LMT features (of different dimensions) lead to significant improvements, raising the performance to more than 70% with 50-dim PCA-reduced LMT features. However, increasing the dimension above 50 degrades the performance, possibly due to overfitting. The tracking performance can significantly impact the recognition results, which can be used as an indirect measure of the tracker's quality. Compared to HMM-based recognition with IVT-tracked face sequences, our visually-constrained tracker leads to nearly 6% improvement in accuracy (HMM LDA+LMT(PCA-50) in Fig. 10). This clearly suggests that feedback from recognition to tracking can be used to learn improved trackers. However, the feedback mechanism needs to be constructed in a manner that will retain generality and scalability of the approach (*i.e.*, identical tracking parameters across different videos/subjects). Recognition is fast, taking just a few seconds for test videos hundreds of frames long.

6. Conclusion

We have addressed the problem of tracking and recognition of faces in real-world noisy videos. Video-based face recognition, though more robust than frame-by-frame recognizers in the presence of illumination and pose variations, requires well-tracked face sequence. Our proposed tracker improves robustness of existing adaptive appearance trackers by introducing additional visual constraints in a particle filtering framework. We have demonstrated that the HMM-based recognizer with hidden states modeled as face poses and LDA-based pose-discriminant features outperforms state-of-the-art recognizers. An extensive set of experiments, including recognition of 35 celebrities in real-world YouTube videos, confirms that the recognition performance can be used as a metric for judging tracker's quality when no ground truth tracking information is available. In the future, we plan to devise a principled way to feed recognition results back into tracking to improve the tracking performance and, in turn, the recognition accuracy.

Acknowledgements

This work was supported in part by NSF IIS grant $\#0413105$. We thank Sergey Ioffe and Harwig Adams for their face detectors and the LMT features, respectively.

References

[1] O. Arandjelovic and R. Cipolla. Face recognition from video using the generic shape-illumination manifold, 2006. ECCV.

[2] S. Avidan. Support vector tracking, 2001. CVPR.

[3] S. Berrani and C. Garcia. Enhancing face recognition from video sequences using robust statistics, 2005. Advanced Video and Signal Based Surveillance (AVSS).

[4] M. J. Black and A. D. Jepson. EigenTracking: Robust matching and tracking of articulated objects using a view-based representation, 1996. ECCV.

[5] R. Chellappa, V. Kruger, and S. Zhou. Probabilistic recognition of human faces from video, 2002. ICIP.

[6] D. Comaniciu, V. Ramesh, and P. Meer. Kernel-based object tracking. *IEEE Trans. on PAMI*, 25(5):564–575, 2003.

[7] D. O. Gorodnichy. Associative neural networks as means for low-resolution video-based recognition, 2005. International Joint Conference on Neural Networks (IJCNN).

[8] G. D. Hager, M. Dewan, and C. V. Stewart. Multiple kernel tracking with SSD, 2004. CVPR.

[9] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density, 1996. ECCV.

[10] A. D. Jepson, D. J. Fleet, and T. F. El-Maraghi. Robust online appearance models for visual tracking. *IEEE Trans. on PAMI*, 25(10):1296–1311, 2001.

[11] M. La Cascia, S. Sclaroff, and V. Athitsos. Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3D models. *IEEE Trans. on PAMI*, 22(4):322–336, 2000.

[12] K.-C. Lee, J. Ho, M. Yang, and D. Kriegman. Visual tracking and recognition using probabilistic appearance manifolds. *Computer Vision and Image Understanding*, 2005.

[13] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds, 2003. CVPR.

[14] Y. Li, H. Ai, T. Yamashita, S. Lao, and M. Kawade. Tracking in low frame rate video: A cascade particle filter with discriminative observers of different lifespans, 2007. CVPR.

[15] Y. Li, S. Gong, and H. Liddell. Video-based online face recognition using identity surfaces, 2001. ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems.

[16] W. Liu, Z. Li, and X. Tang. Spatio-temporal embedding for statistical face recognition from video, 2006. ECCV.

[17] X. Liu and T. Chen. Video-based face recognition using adaptive hidden Markov models, 2003. CVPR.

[18] D. Ross, J. Lim, R.-S. Lin, and M.-H. Yang. Incremental learning for robust visual tracking. *IJCV*, 2007.

[19] J. Stallkamp and R. S. H.K. Ekenel. Video-based face recognition on real-world data, 2007. ICCV.

[20] X. Zhang, W. Hu, S. Maybank, and X. Li. Graph based discriminative learning for robust and efficient object tracking, 2007. ICCV.