# Face Recognition and Retrieval in Video

Caifeng Shan

**Abstract.** Automatic face recognition has long been established as one of the most active research areas in computer vision. Face recognition in unconstrained environments remains challenging for most practical applications. In contrast to traditional still-image based approaches, recently the research focus has shifted towards video-based approaches. Video data provides rich and redundant information, which can be exploited to resolve the inherent ambiguities of image-based recognition like sensitivity to low resolution, pose variations and occlusion, leading to more accurate and robust recognition. Face recognition has also been considered in the content-based video retrieval setup, for example, character-based video search. In this chapter, we review existing research on face recognition and retrieval in video. The relevant techniques are comprehensively surveyed and discussed.

## 1 Introduction

Automatic face recognition has long been established as one of the most active research areas in computer vision [119, 70, 61, 99, 1]. This is mainly due to its wide range of applications such as person identification, law enforcement, smart environment, visual surveillance, human-computer interaction, and image/video retrieval. After three decades of intense research, the state-of-the-art approaches can achieve high recognition rate under controlled settings [86]. However, face recognition in unconstrained real-life environments remains challenging for most practical applications.

Faces are probably the most common cue used by humans for identifying people. Face recognition has mainly been studied for biometric identification. Biometrics refers to the measurement and analysis of physical or behavioral characteristics of humans; various visual traits, such as fingerprint, face, iris, gait and hand geometry,

Caifeng Shan
Philips Research, Eindhoven, The Netherlands
e-mail: caifeng.shan@philips.com

D. Schonfeld et al. (Eds.): Video Search and Mining, SCI 287, pp. 235–260.
springerlink.com

have been explored for biometric recognition [54]. Among different biometric modalities, face recognition allows unobtrusive identification in uncontrolled scenarios without requiring user cooperation, for example, in low quality surveillance footage.

Numerous approaches have been developed for face recognition in the last three decades. Traditionally, face recognition research has mainly focused on recognizing faces from still images [103, 15]. However, single-shot based recognition is hard because of the well-known problems such as illumination change, pose variation, facial expression, and occlusion. The face image differences caused by these factors often exceed those due to identity changes. A single image might not provide enough information for reliable classification. Another problem with image-based face recognition is that it is possible to use a prerecorded face photo to confuse the recognition system to take it as a live subject [101]. Recently, the research focus has shifted more and more towards video-based approaches. The advent of inexpensive video cameras and increased processing power makes it viable to capture, store and analyze face videos. Video inputs provide rich and redundant information in the form of multiple frames, for example, normally in video people show a lot of pose and expression variations. It is widely believed that, by properly extracting the additional information, video-based recognition could resolve the inherent ambiguities of image-based recognition, such as sensitivity to low resolution, pose variations and partial occlusion, leading to more accurate and robust face recognition. Furthermore, video inputs allow to capture facial dynamics that are useful for face identification [60, 83].

Although most of the existing research has been focused on the biometric identification paradigm, recently face recognition has been considered in the content-based video retrieval setup [11, 96]. Face information is important in different kinds of videos, especially in news programs, dramas, and movies. Face recognition could be used for video content description, indexing and mining, e.g., rapid browsing or retrieval of scenes based on the presence of specific actors. Increasing amount of video content on the web is marking a new phase of how users consume information, where users often look for specific people related video content. Current video search engines mainly rely on the keywords that appear in descriptions or in the surrounding web page content. Face recognition enables more accurate video search by focusing on the content of videos.

In this chapter, we review existing research on face recognition in video. The relevant techniques are comprehensively surveyed and discussed. We also introduce recent work on face retrieval in video. The chapter is organized as follows. In Section 2, we briefly introduce face detection and face tracking, the two important components in face recognition and retrieval research. Section 3 discusses video-based face recognition technologies. Specifically, three main categories, including key-frame based, temporal model based, and image-set matching based, are described respectively. In Section 4, we present recent research on face retrieval in video. Challenges and future research directions are discussed in Section 5. Finally Section 6 concludes the chapter.

# 2 Face Detection and Tracking

Face detection and face tracking are important components in face recognition systems, which automatically detect or locate the face region in the input frames. Normally from the located face region, the relevant features are extracted and subsequently served as input to the face recognizer. In this section, we briefly introduce existing work in these two areas.

## *2.1 Face Detection*

Face detection plays a crucial role in face-related vision applications. Due to its practical importance, numerous techniques have been proposed for face detection (see [113] for a survey). In most of existing methods, appearance features such as edge, intensity, and color, are extracted to locate the faces using statistical or geometric models. The real-time face detection scheme proposed by Viola and Jones [104, 105] is arguably the most commonly employed front face detector, which consists of a cascade of classifiers trained by AdaBoost employing Harr-wavelet features. AdaBoost [34, 92] is one of the most successful machine learning techniques applied in computer vision, which provides a simple yet effective approach for stagewise learning of a nonlinear classification function. Later their approach was extended with rotated Harr-like features and different boosting algorithms [76]. In [71], by incorporating Floating Search into AdaBoost, FloatBoost was proposed for improved performance on multi-view face detection.

Many other machine learning techniques such as Neural Network and Support Vector Machine (SVM) have also been introduced for face detection. In [77], the Bayes classifier was adopted with discriminating feature analysis for frontal face detection. The input image, its 1D Harr wavelet representation, and its amplitude projections are combined to derive a discriminating feature vector. Later the features were extended and combined with a SVM-based classifier [94]. SVM was also used with the spectral histogram features for face detection [107]. To improve the detection efficiency, Garcia and Delakis [37] designed a convolutional neural network for face detection, which performs simple convolutional and subsampling operations. More recently, the approach in [77], Viola and Jones's approach [104, 105], and the approach in [37] are modified and combined for a fast and robust face detector [22]. Overall, face detection technology is fairly mature and a number of reliable face detectors have been built based on existing approaches.

## *2.2 Face Tracking*

Most of face detectors can only detect faces in the frontal or near-frontal view. To handle large head motion in video sequences, face or head tracking is usually adopted. In some work [120, 55], face tracking and recognition are integrated in one framework, for example, Zhou *et al.* [122, 120] proposed a probabilistic approach for simultaneous tracking and recognition.

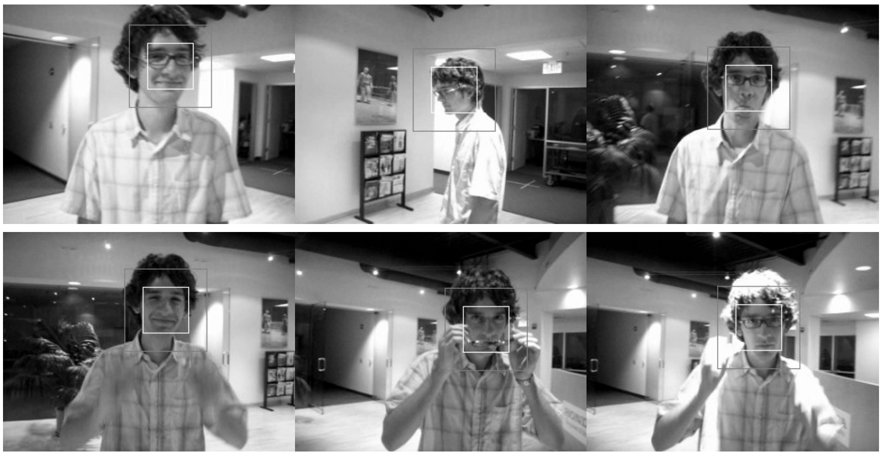

**Fig. 1** Examples of face tracking using the online boosting algorithm [44].

Visual tracking of objects has been intensively studied in computer vision, and different approaches have been introduced, for example, particle filtering [13] and mean shift [19, 23]. Accurate face tracking is difficult because of face appearance variations caused by the non-rigid structure, 3D motion, occlusions, and environmental changes (e.g., illumination). Therefore, adaptation to changing target appearance and scene conditions is a critical property a face tracker should satisfy. Ross *et al.* [88] represented the target in a low-dimensional subspace which is adaptively updated using the images tracked in the previous frames. In [44], Grabner *et al.* introduced the online boosting for tracking, which allows online updating of the discriminative features of the target object. Some face tracking results of their approach are shown in Fig. 1. Compared to the approaches using a fixed target model such as [14], these adaptive trackers are more robust to face appearance changes in video sequences.

One main drawback of these adaptive approaches is their susceptibility to drift, i.e., gradually adapting to non-targets, because the target model is built from the previous tracked results. To address this problem, a mechanisms for detecting or correcting drift should be introduced [55], by adding global constraints on the overall appearance of the target. For faces, such constraints could be learned from a set of generic well-cropped/aligned face images that span possible variations in pose, illumination, and expression. Specifically, two constraint terms were introduced in [55]: (1) a set of facial pose subspaces (or manifolds), each of which represents a particular out-of-plane pose, and (2) a SVM based goodness-of-crop discriminator whose confidence score indicates how well the cropped face is aligned. Grabner *et al.* [45] introduced an online semi-supervised boosting to alleviate the drifting problem. They formulated the update process in a semi-supervised fashion which uses the labeled data as a prior and the data collected during tracking as unlabeled samples.

After the face detection and tracking stages, faces are only roughly localized and aligned. Face registration methods can be adopted to deal with the effect of varying pose, for example, by utilizing the characteristic facial points (normally locations of the mouth and eyes). In some work (e.g., [59]), face recognition is performed directly on faces roughly localized, close to the conditions given by typical surveillance systems.

## 3 Face Recognition in Video

Recent years have witnessed more and more studies on face recognition in video. Zhang and Martinez [114, 115] investigated whether the methods, defined to recognize faces from a single still image, perform better if they could work with multiple images or video sequences. By extending their probabilistic appearance-based approach [80], they showed that regardless of the feature extraction method used, the recognition results improve considerably when using a video sequence rather than a single image. It is also observed in [47] that the spatial-temporal representation derived from video outperforms the image-based counterpart. Video-based recognition provides a setting where weak evidence in individual frames can be integrated over time, potentially leading to more reliable recognition in spite of the difficulties such as pose variation and facial expression.

We partition the existing research into three categories: (1) key-frame (or exemplar) based approaches, (2) temporal model based approaches, and (3) image-set matching based approaches. The first class [47, 115, 98, 101] considers the problem as a recognition problem from still images by independently using all or a subset of the face images. Usually a voting rule is used to come up with a final result. In most of cases, only a subset of representative face images is used for recognition, where ad hoc heuristics are used to select key-frames. The second class [75, 123, 64, 79, 49] makes use of all face images together with their temporal order in the video. By taking into account temporal coherence, face dynamics (such as non-rigid facial expressions and rigid head movements) within the video sequence are modeled and exploited to enforce recognition. The third class [110, 93, 10, 59, 106] also uses all face images, but does not assume temporal coherence between consecutive images; the problem was treated as an image-set matching problem. The distributions of face images in each set are modeled and compared for recognition, and the existing work can be further divided into statistical model-based and mutual subspace-based methods (see Section 3.3 for details). Both the second and third categories integrate the information expressed by all the face images into a single model. The categorization of relevant techniques are summarized in Table 1. In the following sections, we discuss each category in details.

### *3.1 Key-Frame Based Approaches*

The approaches in this category treat each video as a collection of images, and perform face recognition by comparing all or a subset of individual face images

**Table 1** Categorization of video-based face recognition techniques.

| Category | Method |
|---|---|
| Key-frame based Approaches | [90], [40], [47], [114], [100], [17], [115], [31], [78], [85], [98], [101], [118] |
| Temporal Model based Approaches | [74], [73], [72], [75], [18], [24], [67], [69], [68], [122], [120], [123], [121], [64], [65], [66], [79], [55], [2], [43], [50], [49] |
| Image-Set Matching based Approaches | |
| Statistical model-based | [93], [4], [96], [7], [10], [6], [9] |
| Mutual subspace-based | [110], [90], [35], [82], [108], [56], [57], [5], [58], [59], [106], [38] |

with training data using image-based recognition techniques. They neither make any assumption on the underlying distributions of input face images, nor use their temporal coherence. They are based on premise that similarity of the test video with the training data, which could be still images or videos, is reflected by the similarity of the images from the testing video and training data. For example, Satoh [90] matches two face sequences based on face matching between a closest pair of face images across two sequences. These approaches may fail as they do not take into account of the effect of outliers [59]. If requiring a comparison of every pair of samples drawn from the input video and training data, such methods could be time consuming.

Normally a subset of "good" or representative frames (called key-frames or exemplars) is selected to perform image-based recognition. In [40], face recognition is performed based on tracking the nose and eyes. Their locations are used to decide whether the face is suitable for recognition. If they form an equilateral triangle, image-based recognition is performed; otherwise, the tracking continues until an appropriate frame occurs. Berrani and Garcia [17] proposed to select good-quality face images using robust statistics. Specifically, by considering it as an outlier detection problem, they utilized RobPCA to filter out the noisy face images (e.g., not well-centered, non-frontal). Their experiments on two face image databases show that the filtering procedure improves the recognition rate by 10% to 20%. In [85], three face matchers are fused for face recognition in video, where the estimated face poses and the detected motion blur are used for adaptive fusion, e.g., frames with motion blur are not considered for recognition. Experimental results on the CMU Face-In-Action database [39] with 204 subjects show that their approach achieves consistent improvements.

It is argued in [63] that the best exemplars are those which minimize the expected distance between the video frames and the exemplars; radial basis functions are applied to select exemplars. Hadid and Pietikäinen [47] proposed to extract the most representative faces by applying K-Means clustering in the low-dimensional space derived by Locally Linear Embedding [89]. They adopted a probabilistic voting to combine image-based recognition over video frames for final decision. In [31], following the Isomap algorithm [102], the geodesic distances between face images are estimated, based on which a hierarchical agglomerative clustering algorithm is applied to derive local face clusters of each individual. The authors argued that using a single exemplar for each cluster may not fully characterize the image variability, and proposed to construct two subspaces to characterize intra-personal and extra-personal variations for each local cluster. Given a test image, the angle between its projections onto the two subspace is used as a distance measure to the cluster. Experiments on a video data set of 40 subjects demonstrate their approach produces promising results compared to previous methods. Zhao and Yagnik [118] presented an approach for large scale face recognition in web videos. For each face set derived by facial feature tracking, key faces are selected by clustering; the face sets are further clustered to get more representative key faces and remove duplicate key faces. A combination of majority voting and probabilistic voting is adopted for final decision. Evaluated on large-scale web videos, their approach achieves 80% top-5-precision on tested persons.

Tang and Li [100, 101] proposed to align video sequences of different subjects based on the audio signal in video, i.e., frames with similar facial expressions are synchronized according to the associated speech signal. A number of distinctive frames are selected from the synchronized video sequences. Key fiducial points on each face image are further aligned, and Gabor wavelet features are extracted from these points for facial representation. For matching the spatial and temporal synchronized video sequences, they developed a multi-level discriminant subspace analysis algorithm. They also integrated the frame-based classification using the majority voting or sum rule. In [78], Liu *et al.* proposed a synchronized frame clustering method which incrementally outputs aligned clusters across all video sequences, and adopted a Bayesian method to select key-frames. A Nonparametric Discriminant Embedding is introduced for learning spatial embedding. With the spatial-temporal embedding of video sequences, they presented a statistical classification solution, which uses a probabilistic voting strategy to combine the recognition confidences in each frame. Encouraging results on the XM2VTS database [81] are reported in these studies [100, 101, 78].

Stallkamp *et al.* [98] presented a real-time video-based face recognition system which recognizes people entering through the door of a laboratory. As shown in Fig. 2, without user cooperation, the captured video data contains difficult situations arising from pose variations, facial expressions, occlusions due to accessories and hair, illumination changes due to the time and weather conditions and light switched on/off. Their approach combines the individual frame-based classification results to one score per sequence. With DCT-based appearance features, individual frame scores are generated using a k-nearest neighbor classifier and a set of Gaussian mixture models learned from training sequences. They proposed two measures to weight the contribution of each individual frame: *distance-to-model* (DTM) and *distance-to-second-closest* (DT2ND). DTM takes into account how similar a test sample is to the representatives of the training set. Test samples that are very different from the training data are more likely to cause a misclassification, so DTM is used to reduce the impact of these samples on the final score. Based on the idea that

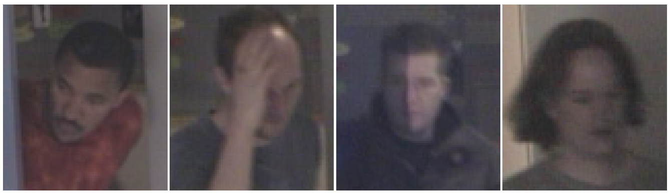

**Fig. 2** The real-world video data used in [98], which shows a variety of different lighting, pose and occlusion conditions.

reliable matching requires the best match to be significantly better than the second-best match, DT2ND is used to reduce the impact of frames which deliver ambiguous classification results. Their experiments on a database of 41 subjects show both measures have positive effects on the classification results. Despite promising results, the need for parameter tuning and heuristic integration schemes may limit the generalization of this approach.

### *3.2 Temporal Model Based Approaches*

Other than the multitude of still frames, video allows to characterize faces based on the inherent dynamics which is not possible with still images. Facial dynamics include the non-rigid movement of facial features (e.g., facial expressions) and the rigid movement of the whole face (e.g., head movements). Psychological studies [60, 83, 95] indicate that facial dynamics play an important role in the face recognition process, and both static and dynamic facial information are used in the human visual system to identify faces. Facial dynamics are even more crucial under degraded viewing conditions such as poor illumination, low resolution, and recognition at distance. Many of these points have been verified in computer vision research [48]. For example, Gorodnichy [41, 42] illustrated the lack of resolution can be compensated by the dynamic information coming from the time dimension. Many approaches have been proposed to utilize the temporal continuity inherent in videos for face recognition [20].

Li *et al.* [74, 73, 72, 75] proposed to model facial dynamics by constructing facial identity structures across views and over time, referred to identity surfaces (shown in Fig. 3), in the Kernel Discriminant Analysis feature space. Dynamic face recognition is performed by matching the face trajectory computed from a video input and a set of model trajectories constructed on the identity surfaces. The trajectory encodes the spatio-temporal dynamics of moving faces, while the trajectory distance accumulates recognition evidence over time. Experimental results on video sequences of 12 subjects were reported with a recognition rate of 93.9%. Similarly, in [18], each image sequence of a rotating face is projected into the eigen-space using Principal Component Analysis (PCA) and represented as a trajectory in the space; face recognition is performed as the trajectory matching. They reported excellent recognition rates on a data set of 28 subjects.

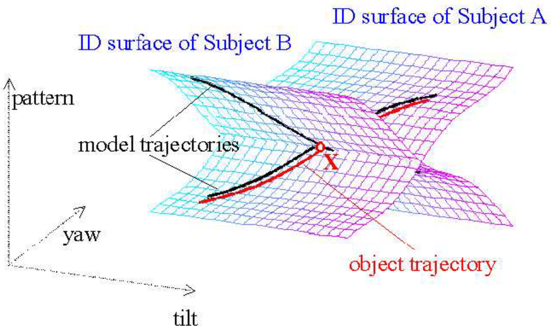

**Fig. 3** Identity surfaces for dynamic face recognition [73].

Edwards *et al.* [24] learnt how individual faces vary through video sequences by decoupling sources of image variations such as expression, lighting and pose. The trained statistical face model is used to integrate identity evidence over a sequence, which is more stable and robust than a model trained from a single image. Li and Chellappa [67, 69] presented an approach to simultaneous tracking and verification in video, based on posterior density estimation using sequential Monte Carlo methods; verification is realized through hypothesis testing using the estimated posterior density. By rectifying each face template onto the first frame of the testing video, the approach has been applied to face verification in video. They [68] also introduced a method for face verification using the motion parameters of tracked facial features, where the features are extracted using Gabor filters on a regular 2D grid. Their method produces encouraging results on a data set with 19 subjects.

Following [67, 69], Zhou *et al.* [122, 120] proposed a probabilistic approach for simultaneous tracking and recognition in video. They used a time-series state space model which is parameterized by a tracking state vector (continuous) and a identity variable (discrete), in order to simultaneously characterize the evolving kinematics and identity. The joint posterior probability is approximated and propagated using the Sequential Importance Sampling algorithms, and the marginal distribution of the identity variable is estimated to provide the identity result. In the still-to-video setting, where the gallery consists of still images and the probe consists of videos, the approach was evaluated on two data sets with 12 subjects and 30 subjects respectively. In [123], to consider the video-to-video setting (i.e., generalizing the gallery to videos), they adopted an exemplar learning method [63] to select representatives from the gallery videos, serving as still templates in the still-to-video scenario. Their approach was tested on the MoBo database [46]. Later the simultaneous tracking and recognition approach was improved by incorporating appearance-adaptive models [121]. The appearance changes between input frames and gallery images are modeled by constructing the intra- and extra-personal spaces. Experiments on a data set of 29 subjects illustrate the approach can handle appearance changes caused by pose and illumination variations, leading to improved tracking and recognition performance.

To address continues head pose changes in video, Lee *et al.* [64] proposed to learn a low-dimensional appearance manifold of faces, which is approximated by piecewise linear subspaces (named pose manifolds). To construct the representation, exemplars are first sampled from videos by finding frames with the largest distance to each other, which are further clustered using K-means clustering. Each cluster models face appearance in nearby poses, represented by a linear subspace computed by PCA. The dynamics among pose manifolds are encoded as transition probabilities, learned from training video sequences. They presented a maximum a posteriori formulation for face recognition, which integrates the likelihood that the input image comes from a particular pose manifold and the transition probability to this pose manifold from the previous frame. Their approach was extended for simultaneously tracking and recognizing faces in video [65], achieving the recognition rate of 98.8% on a data set of 20 subjects. However, the appearance model in these works was learned by a batch training process from short video clips, which is not practical for large number of lengthy video sequences. In [66], an online learning algorithm was introduced to incrementally construct a person-specific appearance manifold using an initial generic prior and successive frames from the video of a subject. Experimental results demonstrate the approach constructs an effective representation for face tracking and recognition.

Liu and Chen [79] introduced to use the adaptive Hidden Markov Model (HMM) for video-based face recognition. In the training phase, a HMM is created for each individual to learn the statistics and temporal dynamics using the eigen-face image sequence. During the recognition process, the test sequence is analyzed over time by the HMM corresponding to each subject; its identity is determined by the model providing the highest likelihood. In case of face recognition with low-quality images, the HMM-based method was shown to produce better results than image-based methods [47]. Considering PCA features may not be sufficiently discriminative among multiple head poses, Kim *et al.* [55] proposed to use Linear Discriminant Analysis (LDA) coupled with the modeled pose dynamics in the HMM framework. By fusing pose-discriminant and person-discriminant features over the video sequence, their approach leads to superior performance, e.g., recognition rate of over 70% on a YouTube video set containing 35 celebrities in 1500 video sequences.

Aggarwal *et al.* [2] posed video-based face recognition as a dynamical system identification problem. A moving face is modeled as a linear dynamical system, and each frame is regarded as the output of the system. They adopted an autoregressive and moving average (ARMA) model to represent such a system. The similarity between gallery and probe video sequences is computed using principal angle based metrics. Their approach performs well on the data set of 16 subjects and the UCSD/Honda database [64]. Gorodnichy [43] proposed to use the neuro-associative principle for face recognition, according to which both memorization and recognition are done based on a flow of frames. The temporal dependence between consecutive images is considered by adding extra neurons. This approach achieves

**Fig. 4** Example frames from video sequences in the IIT-NRC database.

recognition rate of over 95% on the IIT-NRC database[1] of 11 subjects. Some example frames of this database are shown in Fig. 4.

Recently texture-based spatiotemporal representations have been exploited for analyzing faces in video. In [117], volume Local Binary Patterns (LBP) based description was applied to facial expression recognition. Hadid *et al.* [50, 49] proposed to use local volumetric spatio-temporal features for face recognition, by considering a face sequence as a selected set of volumes from which local histograms of extended volume LBP are extracted. They adopted the boosting scheme to learn the discriminative local volumetric features, where all combination of sequence pairs are considered as the intra- and extra-classes. Experimental results on three public databases demonstrate their approach provides superior recognition performance. We compare the reported recognition performance on several public datasets in Table 2.

### 3.3 Image-Set Matching Based Approaches

While in some cases temporal dynamics could be exploited, in a more general scenario, the extracted face images may not be temporally consecutive. This could be due to the practical limitations in current face acquisition process, e.g., it is difficult to continuously detect or track face from every video frame, or the images are the sparse and unordered observations collected over an long periods of time and from multiple viewpoints. It is then not possible to model facial dynamics for face recognition. Image-set matching based approaches formulate face recognition as matching a probe image set against all the gallery image sets each of which representing one subject, without assuming temporal coherence between consecutive images. They can be applied to image sets containing ordered observations collected over consecutive time steps, and also image sets acquired by multiple independent still shots (e.g., photo collections) or long-term discontinuous observations.

In [59], relevant approaches to image set classification are divided into two categories: non-parametric sample-based and parametric model-based. Non-parametric sample-based approaches are based on matching pair-wise samples in the image

[1] http://synapse.vit.iit.nrc.ca/db/video/faces/cvglab.

**Table 2** Recognition results on several public databases in the literature.

| Database | Subjects/Sequences | Approach | Result(%) |
|---|---|---|---|
| Honda/UCSD[64] | 20/40 | Probabilistic appearance manifold [64] | 93.2 |
| | | Exemplar-based probabilistic voting [47] | 86.5 |
| | | System identification[2] | 90.0 |
| | | Extended probabilistic appearance manifold [65] | 98.8 |
| | | HMM (LDA+LandMark Template) [55] | 100 |
| | | Boosted extended volume LBP [49] | 96.0 |
| | | Manifold-Manifold Distance [106] | 96.9 |
| MoBo[46] | 24/96 | Adaptive HMM [79] | 98.8 |
| | | Exemplar-based probabilistic voting [47] | 90.8 |
| | | Boosted extended volume LBP [49] | 97.9 |
| | | Manifold-Manifold Distance [106] | 93.6 |
| XM2VTS[81] | 295/1180 | Multi-level discriminant subspace analysis [101] | 99.3 |
| | | Multi-classifier (voting or sum) [101] | 99.3 |
| | | Spatial-temporal embedding [78] | 99.3 |

sets. To recognize the two image sets as the same class, a solution would be to find the representative images from each image set and then measure their similarity. These approaches have been discussed in Section 3.1. Parametric model-based approaches tend to represent each image set by a parametric distribution and then measure the similarity between two distributions [93, 7, 10]. This is based on the assumption that images are drawn from some distributions on the underlying face pattern manifold, and normally statistical learning algorithms are adopted to model the distribution. Recently, following the mutual subspace method [110], many approaches build a compact model of the distribution by representing each image set as a linear subspace, and measure their similarity using the canonical angles [110, 82, 59]. In the following sections, we discuss these two groups of approaches: statistical model-based and mutual subspace-based, respectively.

### 3.3.1 Statistical Model-Based Approaches

Shakhnarovich *et al.* [93] cast face recognition from an image set as a statistical hypothesis testing problem, with the assumption that images are independently and identically (i.i.d) drawn samples from a probability density function (pdf). They proposed to classify sets of face images by comparing the probability distributions of the probe set and the gallery sets. Specially, they estimated the face appearance distribution by a multivariate Gaussian, and used the Kullback-Leibler (KL) divergence, which quantifies how well a particular pdf describes samples from another pdf, to measure the similarity. Evaluation on two data sets of frontal face images, with 29 subjects and 23 subjects respectively, demonstrates their approach achieves equal or improved recognition performance compared to image-based recognition methods and the mutual subspace method. However, to make the divergence computation tractable, they made a crude assumption that face patterns are normally distributed, which may not be true [108]. Arandjelovic and Cipolla [4, 7] argued against the use of KL divergence due to its asymmetry, and proposed to use the Resistor-Average Distance (RAD) as the dissimilarity measure between distributions. They adopted kernel PCA to solve the closed-form expression of RAD between two Gaussian distributions, which also allows for expressive modeling of nonlinear face manifolds. In addition, a data-driven approach was used for stochastic manifold repopulating, in order to generalize unseen modes of variation. Their approach achieves recognition rate of 98% on a database of 100 individuals collected under varying imaging conditions, outperforming KL divergence based approaches and the mutual subspace method.

To deal with nonlinear variations in face appearance due to illumination and pose changes, Arandjelovic *et al.* [10] model the face appearance distribution as Gaussian Mixture Models (GMMs) on low-dimensional manifolds. The KL divergence was adopted to measure the similarity between the estimated distributions. The advantage of this approach over the previous kernel method [4, 7] lies in its better modeling of distributions confined to nonlinear manifolds; however this benefit comes at the cost of increased difficulty of divergence computation. The KL divergence, which for GMMs cannot be computed in the closed form, is evaluated by a Monte-Carlo algorithm. They evaluated the proposed method on a data set with 100 subjects, and obtained the average performance of 94%. In [6], they derived a local manifold illumination invariant, and formulated the face appearance distribution as a collection of Gaussian distributions corresponding to clusters obtained by K-means. The image set matching was performed by pair-wisely comparing all clusters of two manifolds and the maximal of cluster similarities is chosen as the overall manifold similarity. To compare two Gaussian clusters, they proposed to find the most probable mode of mutual variations between the two clusters. Recently they [9] proposed to decompose each face appearance manifold into three Gaussian pose clusters describing small face motion around different head poses. Given two manifolds, the corresponding pose clusters are compared, and the pair-wise comparisons are combined to measure the similarity between manifolds. To achieve illumination invariant recognition, they considered coarse region-based Gamma correlation with fine illumination manifold-based normalization. Their approach demonstrated consistently superior recognition performance on a database with 60 individuals.

Statistical model-based approaches make strong assumptions about the underlying distributions. The main drawbacks are that they need to solve the difficult parameter estimation problem and they easily fail when the training sets and the testing sets have weak statistical relationships, for example, when they are from different ranges of poses, expressions or illumination changes.

### 3.3.2 Mutual Subspace-Based Approaches

The distribution of a set of face images can be compactly represented by a lower-dimensional linear subspace. Yamaguchi *et al.* [110] introduced the Mutual Subspace Method (MSM), where each image set is represented by the linear subspace spanned by the principal components of the images. The similarity between image sets is measured by the smallest principal angles between subspaces. Principal angles [51] are the minimal angles between vectors of two subspaces, which reflect the common modes of variation of two subspaces. Canonical correlations, which are cosines of principal angles, are often used as the similarity measure. Later, to make it insensitive to variations such as pose and illumination changes, MSM was extended to the Constrained Mutual Subspace Method (CMSM) [35]. In CMSM, the test subspace and the reference subspace are projected onto a constraint subspace, where each subspace exhibits small variance and the two subspaces could be better separated. A real-time system implemented using CMSM was demonstrated in [62]. In [82], the authors further introduced the Multiple Constrained Mutual Subspace Method (MCMSM), which generates multiple constraint subspaces by using the ensemble learning algorithms (Bagging and Boosting). The similarities on each constraint subspace are combined for recognition. They conducted experiments on a database of 50 subjects and a database of 500 subjects, and experimental results show improved recognition performance compared to MSM and CMSM.

An attractive feature of MSM-based methods is their computational efficiency [56]: principal angles between linear subspaces can be computed rapidly, while the estimation of linear subspaces can be performed in an incremental manner. However, the simplistic modeling using a linear subspace cannot sufficiently model complex and nonlinear face appearance variations, and is sensitive to particular data variations. It is also argued in [93, 106] that MSM-based methods could not consider the entire probabilistic model of face variations, since the eigenvalues corresponding to the principal components, as well as the means of the samples, are disregarded.

There have been some attempts to extend MSM-based methods for nonlinear subspaces or manifolds [108, 56, 106]. Wolf and Shashua [108] introduced to compute principal angles between nonlinear manifolds using the kernel trick. However, as in all kernel approaches, finding the optimal kernel function is a difficult problem. Kim *et al.* [56, 57] argued that MSM-based methods have two shortcomings: the limited capability of modeling nonlinear pattern variations and the ad hoc fusion of information contained in different principal angles. They extended the concept of principal angles to nonlinear manifolds by combining global manifold variations with local variations, where the locally linear manifold patches are obtained using mixtures of Probabilistic PCA. The similarity between manifolds is computed as a weighted average of the similarity between global modes of data variation and the best matching local patches. They further adopted AdaBoost to learn the application-optimal principal angle fusion. Experiments on a database with 100 subjects demonstrate the above two nonlinear manifold modeling approaches both achieve superior performance to the basic MSM. By decomposing the nonlinear manifold as a collection of local linear models, each depicted by a linear subspace, Wang *et al.* [106]

introduced to compute the Manifold-Manifold Distance, which is defined as the distance of the closest subspace pair from two manifolds. Regarding the subspace-subspace distance, they argued principal angles mainly reflect the common modes of variation between two subspaces while ignoring the data itself, so they proposed to also consider the sample means in the local models to measure the local model similarity. Experiments on two public databases demonstrate their approach produces superior performance to the MSM method.

Using canonical correlations as the distance measure of image sets, Kim *et al.* [58, 59] proposed a discriminative learning method for image set classification. They developed a linear discriminant function that maximizes canonical correlations of within-class sets and minimizes canonical correlations of between-class sets. Image sets transformed by the discriminant function are then compared by the canonical correlations. There approach was evaluated on various object recognition tasks, achieving consistently superior recognition performance. In [38], a loss based Regularized LDA is introduced for face image set classification using canonical correlations.

## 4 Face Retrieval in Video

Face recognition could be used for video content description, indexing, and retrieval [25, 96]. The dramatic increase of videos demands more efficient and accurate access to video content. Finding a specific person in videos is essential to understand and retrieve videos [91]. Face information is an important cue for person identification in many types of videos such as news programs, dramas, movies, and homemade videos. Face retrieval enables many new functionality, e.g., rapid browsing, where only shots containing a specific character are chosen to play. Face recognition also allows character-based video search, which receives growing interest due to huge amount of video content online (e.g., YouTube).

Face retrieval in general is to retrieve shots containing particular persons/actors given one or more query face images. In context of videos captured in real-life scenarios (e.g., news, programs, and films), lighting variations, scale changes, facial expressions and varying head pose drastically change the appearance of faces. Partial occlusions, because of the objects in front of faces or resulting from hair style changes also cause problems. Artefacts caused by motion blur and low resolution are also common. In brief, the uncontrolled imaging conditions makes face retrieval very challenging [5]. Some example faces in films are shown in Fig. 5. The existing studies mainly address two kinds of applications: person retrieval and cast listing. In this section, we review relevant approaches for these applications.

**Person Retrieval —** Everingham and Zisserman [29, 28, 30] addressed finding particular characters in situation comedies, given a number of labeled faces (for each character) as training data. They [29] used a skin color model with multi-scale blob detection to detect candidate face regions in the input frame. To deal with pose variations, a coarse 3D ellipsoid head model with multiple texture maps was used to render faces from the train data at the same pose as the target face. The identity of

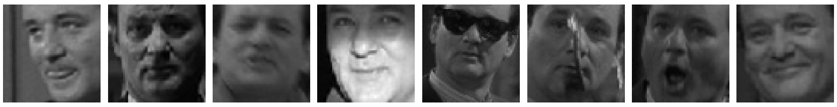

**Fig. 5** Faces in films exhibits a great variability depending on the extrinsic imaging conditions.

the target face is determined by comparing it with the rendered views. The texture maps of the model can be automatically updated as new poses and expressions are detected. In [30], rendered images are used to train a discriminative tree-structured classifier, which detects the individual and estimates the pose over a range of scale and pose. The identity is verified using a generative approach. Their approach was evaluated on 4,400 key-frames (1,500 key-frames in [29]) from a TV situation comedy for detecting three characters. They [28] proposed to synthesize additional training data from a single training image by fitting the 3D model to the person's head. The parts-based constellation models are trained, which propose candidate detections in the input frame. The 3D model is aligned to the detected parts, and the global appearance is considered for recognition. Sivic *et al.* [96] presented a video shot retrieval system based on faces in the shot. Instead of matching single faces, they proposed to match sets of faces for each person, where sets of faces (called face-tracks) are gathered automatically in shots by tracking. Each face in the face-track is described by a collection of five SIFT descriptors around salient facial features. The entire face-track is represented as a distribution (i.e., histogram) over vector quantized face exemplars, resulting in a single feature vector. Face-tracks are matched by comparing the vectors using the chi-square statistics.

Arandjelovic and Zisserman [11, 12] built a system to retrieve film shots based on the presence of specific characters, given one or multiple query face images. To address the variations on scale, pose, and illumination, as well as occlusion, encountered in films, they proposed to obtain a *signature image* by a cascade of processing steps for each detected face. In the first step, facial feature are detected by SVMs that are trained on image patches surrounding eyes and mouth. Face images are then affine warped to have salient facial features aligned. Considering the bounding box typically contains background clutter, the face is segmented from the surrounding background using the face outline, which is detected by combining a learnt prior on the face shape and a set of measurements of intensity discontinuity. Finally illumination effects are removed by band-pass filtering. The end signature image is insensitive to illumination changes, pose variations, and background clutter, and mainly depends on the person's identity (and expression). Signature images are matched using a distance measure for person retrieval. Evaluations on several feature-length films demonstrate that their system achieves recall rates (over 92%) whilst maintaining good precision (over 93%). In [90], Satoh presented comparative evaluation of face sequence matching methods in drama videos.

Face information has been combined with text information for face or person retrieval in video [21, 111, 53, 112, 84]. In [111], multimodal content in videos, including names occurred in the transcript, face information, anchor scenes, and the timing pattern between names and appearances of people, are exploited to find specific persons in broadcast news videos. However, face information was given very small weight in their system. Ozkan and Duygulu [84] also integrated text and face information for person retrieval. Specifically, they limit the search space for a query name by choosing the shots around which the name appears. To find the most-frequently occurring faces in the space, they construct a graph with nodes corresponding to all faces in the search space, and edges corresponding to the similarity of the faces; the problem is transformed into finding the densest component in the graph. A limitation of their approach is that it cannot find a person in parts of the video where his/her name is not mentioned. Zhao and Yagnik [118] presented a large scale system that can automatically learn and recognize faces in web videos by combining text, image, and video. To address the difficulty of manually labeling training data for a large set of people, they used the text-image co-occurrence in the web as a weak signal of relevance, and proposed consistence learning to learn the set of face models from the very large and nosy training set.

**Cast Listing —** An interesting face retrieval problem is automatically determining the cast of a feature-length film. Cast listing has been mainly based on the recognition of faces, as faces being the most repeatable cue in this setting [5], although others cues, such as clothes, can be used as additional evidence. It is challenging because the cast size is unknown, with great face appearance changes caused by extrinsic imaging factors (e.g., illumination, pose, expression). Fitzgibbon and Zisserman [33] made an earlier attempt to this problem using affine invariant clustering. They developed a distance metric that is invariant to affine transformations. Two classes of priors were considered in the distance metric: deformation priors between any pair of frames, and temporal coherence prior between continuous frames. To address the lighting variations, they utilized a simple bandpass filter to pre-process the detected faces. Their approach was tested on the film ’Groundhog Day’ with a principal cast of around 20. Arandjelovic and Cipolla [5] introduced an approach based on clustering over face appearance manifolds, which correspond to sequences of moving faces in a film. They temporally segment the video into shots, and obtain face tracks by connecting face detections through time. The CMSM method [35] is adopted for pair-wise comparisons of face tracks (i.e., face manifolds). To allow unsupervised discriminative learning on an unlabeled set of video sequences, their approach starts from a generic discriminative manifold and converges to a data-specific one, automatically collecting within-class data. Evaluation on a situation comedy illustrates the effectiveness of their method. However, it remains challenging to obtain a small number of clusters per character without merging multiple characters into a single cluster. In [36], normalized Graph Cuts is adopted to cluster face tracks for cast indexing.

In the above cast listing systems, all faces of a particular character are collected into one or a few clusters, which are then assigned a name manually. Everingham *et al.* [26, 27] addressed the problem of automatically labeling faces of characters in TV or film materials with their names. Similar to the “Faces in the News” labeling in [16], where detected frontal faces in news images are tagged with names appearing in the news story text, they proposed to combine visual cues (face and cloth) and textual cues (subtitle and transcript) for assigning names. Regarding face processing [3], face detections in each frame are linked to derive face tracks, and each face is represented by local appearance descriptors computed around 13 facial features. Clothing appearance was also considered as additional cues. They align the transcripts with subtitles using dynamic time warping to obtain textual annotation, and use visual speaker detection to resolve the ambiguities, i.e., only associating names with face tracks where the face is detected as speaking. A nearest neighbor [26] or SVM [27] classifier, trained on labeled tracks, is used to classify the unlabeled face tracks. Their approach has demonstrated promising performance on three 40 minute episodes of a TV serial. In [97], the approach was further extended for improved coverage by character detection and recognition in profile views. Considering there are not enough temporally local name cues in the subtitle and script for local face-name matching, Zhang *et al.* [116] proposed to perform a global matching between the clustered face tracks and the names extracted from the film script. A graph matching method is utilized to build face-name association between a face affinity network and a name affinity network. Experiments on ten films demonstrate encouraging results.

Ramanan *et al.* [87] introduced a semi-supervised method for building large labeled face datasets by leveraging archival video. They implemented a system for labeling archival footage spanning 11 years from the television show Friends. The dataset they compiled consists of more than 600,000 faces, containing appearance variations due to age, weight gain, changes in hairstyles, and other factors. In their system, at the lowest level, detected frontal faces are clustered and tracked using the color histogram of the face, hair and torso. By part-based color tracking, faces with extreme pose changes are also collected in the clusters. The resulting face tracks are reliable since body appearance is stable over short time scales. At the scene level, face tracks are grouped by an agglomerative clustering procedure based on body appearance, since people tend to ware consistent clothes within a scene. They manually labeled the clusters from a single reference episode, which are used to label the dataset using a nearest-neighbors framework.

## 5 Challenges and Future Directions

With camera sensors become pervasive in our society, video data has been one of the main sensory inputs in electronic systems. Meantime, huge amounts of video content have been generated. Face recognition in video, which has many applications such as biometric identification, video search and retrieval, visual surveillance, and human-computer interaction, has received much interest in the last decade. Although much progress has been made, the problem remains difficult for videos captured in unconstrained settings. Some major challenges that should be addressed in future research are considered here:

- **Databases:** Most of the existing public datasets [81, 46, 64] were collected under controlled (laboratory) conditions, with limited number of subjects, covering limited face appearance variations. It is believed that databases built from "the wild" are important for training and evaluating real-world recognition systems. Not only does lack of the large realistic database prevent suitable evaluation of the existing techniques, but also provides little encouragement to novel ideas. Recently the "Labeled Faces in the Wild" database [52] has been collected for unconstrained face recognition. The database contains more than 13,000 labeled face photographs collected from the web, spanning the range of conditions encountered in real life. However, only frontal face images are included in this database. We believe a large database consisting of face videos "in the wild" is necessary for future research. How to build a comprehensive face video database with low manual effort should be investigated [87].

- **Low-quality Video Data:** In many real-life applications, the video data is of low quality (e.g., limited imaging resolution or low frame rate), such as video footage from surveillance cameras and videos captured by consumers via mobile or wearable cameras. The low-quality data could be caused by the poor sensor performance, motion blur, environmental factors such as illumination, or video compression. The sensors in the non-visible spectrum (e.g. Near-Infrared) also generate low-resolution videos with much noise. Existing face recognition approaches mainly focuses on good-quality video data, which cannot be directly applied to low-quality video data. To enable practical applications, it is necessary to investigate face recognition techniques for low-quality video data. Super-resolution could be a solution [8], and temporal information in the video could be used to compensate for the lost spatial information.
- **Computational Cost:** Many applications of face recognition can be foreseen on platforms with limited computing power, for example, video retrieval in mobile devices, smart cameras for video surveillance, user interface of consumer electronics (e.g., toy robotics). The processing power on these devices is limited for traditional video processing tasks like face recognition. Although advanced hardware design and algorithm optimization could be helpful to certain extent [32], most of existing video-based face recognition approaches require high computation, which prevents them for wide applications in low-performance systems. Therefore, there is a great need to investigate low-cost algorithms for face recognition in video.

## 6 Conclusions

Face recognition in video has been an active topic in computer vision, due to many potential applications. This chapter brings a comprehensive review on existing research in this area. Different types of approaches to this problem are discussed. We also introduce recent work on face retrieval. Finally, some challenges and research directions are discussed. Although we mainly focus on video-based approaches, recent years have witnessed some interesting still-image based approaches (e.g., [109]), which could be helpful for face recognition in video.

## References

1. Abate, A.F., Nappi, M., Riccio, D., Sabatino, G.: 2d and 3d face recognition: A survey. Pattern Recognition Letters 28(14), 1885–1906 (2007)
2. Aggarwal, G., Roy-Chowdhury, A.K., Chellappa, R.: A system identification approach for video-based face recognition. In: International Conference on Pattern Recognition (ICPR), pp. 175–178 (2004)
3. Apostoloff, N., Zisserman, A.: Who are you? - real-time person identification. In: British Machine Vision Conference (BMVC), pp. 509–518 (2007)
4. Arandjelović, O., Cipolla, R.: Face recognition from image sets using robust kernel resistor-average distance. In: International Workshop on Face Processing in Video, p. 88 (2004)
5. Arandjelović, O., Cipolla, R.: Automatic cast listing in feature-length films with anisotropic manifold space. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 1513–1520 (2006)
6. Arandjelović, O., Cipolla, R.: Face set classification using maximally probable mutual modes. In: International Conference on Pattern Recognition (ICPR), vol. 1, pp. 511–514 (2006)
7. Arandjelović, O., Cipolla, R.: An information-theoretic approach to face recognition from face motion manifolds. Image and Vision Computing 24(6), 639–647 (2006)
8. Arandjelović, O., Cipolla, R.: A manifold approach to face recognition from low quality video across illumination and pose using implicit super-resolution. In: IEEE International Conference on Computer Vision, ICCV (2007)
9. Arandjelović, O., Cipolla, R.: A pose-wise linear illumination manifold model for face recognition using video. Computer Vision and Image Understanding 113(1), 113–125 (2009)
10. Arandjelović, O., Shakhnarovich, G., Fisher, G., Cipolla, J.R., Zisserman, R.,, A.: Face recognition with image sets using manifold density divergence. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 581–588 (2005)
11. Arandjelović, O., Zisserman, A.: Automatic face recognition for film character retrieval in feature-length films. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 860–867 (2005)
12. Arandjelović, O., Zisserman, A.: On film character retrieval in feature-length films. In: Hammoud, R. (ed.) Interactive Video: Algorithms and Technologies. Springer, Heidelberg (2006)
13. Arulampalam, M., Maskell, S., Gordon, N., Clapp, T.: A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. IEEE Transactions on Signal Processing 50(2), 174–189 (2002)
14. Avidan, S.: Support vector tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(8), 1064–1072 (2004)
15. Belhumeur, P.N., Hespanha, J.P., Kriegman, D.J.: Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. IEEE Transactions on Pattern Analysis and Machine Intelligence 19(7), 711–720 (1997)
16. Berg, T.L., Berg, A.C., Edwards, J., Maire, M., White, R., Teh, Y.W., Learned-Miller, E., Forsyth, D.A.: Names and faces in the news. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 848–854 (2004)
17. Berrani, S.A., Garcia, C.: Enhancing face recognition from video sequences using robust statistics. In: IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS), pp. 324–329 (2005)

18. Biuk, Z., Loncaric, S.: Face recognition from multi-pose image sequence. In: International Symposium on Image and Signal Processing and Analysis, pp. 319–324 (2001)
19. Bradski, G.: Computer vision face tracking for use in a perceptual user interface. Intel Technology Journal (Q2) (1998)
20. Chellappa, R., Aggarwal, G.: Video biometrics. In: International Conference on Image Analysis and Processing, pp. 363–370 (2007)
21. Chen, M.Y., Hauptmann, A.: Searching for a specific person in broadcast news video. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 1036–1039 (2004)
22. Chen, Y.N., Han, C.C., Wang, C.T., Jeng, B.S., Fan, K.C.: A cnn-based face detector with a simple feature map and a coarse-to-fine classifier. IEEE Transactions on Pattern Analysis and Machine Intelligence (2010)
23. Comaniciu, D., Ramesh, V., Meer, P.: Real-time tracking of non-rigid objects using mean shift. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 142–149 (2000)
24. Edwards, G., Taylor, C., Cootes, T.: Improving identification performance by integrating evidence from sequences. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1486–1491 (1999)
25. Eickeler, S., Wallhoff, F., Lurgel, U., Rigoll, G.: Content based indexing of images and video using face detection and recognition methods. In: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), pp. 1505–1508 (2001)
26. Everingham, M., Sivic, J., Zisserman, A.: "hello! my name is.. buffy" automatic naming of characters in tv video. In: British Machine Vision Conference (BMVC), pp. 889–908 (2006)
27. Everingham, M., Sivic, J., Zisserman, A.: Taking the bite out of automated naming of characters in tv video. Image and Vision Computing 27(5), 545–559 (2009)
28. Everingham, M., Zisserman, A.: Automated visual identification of characters in situation comedies. In: International Conference on Pattern Recognition (ICPR), pp. 983–986 (2004)
29. Everingham, M., Zisserman, A.: Automatic person identification in video. In: International Conference on Image and Video Retrieval (CIVR), pp. 289–298 (2004)
30. Everingham, M., Zisserman, A.: Identifying individuals in video by combining 'generative' and discriminative head models. In: IEEE International Conference on Computer Vision (ICCV), vol. 2, pp. 1103–1110 (2005)
31. Fan, W., Yeung, D.Y.: Locally linear models on face appearance manifolds with application to dual-subspace based classification. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, pp. 1384–1390 (2006)
32. Farrugia, N., Mamalet, F., Roux, S., Yang, F., Paindavoine, M.: Fast and robust face detection on a parallel optimized architecture implemented on fpga. IEEE Transactions on Circuits and Systems for Video Technology 19(4), 597–602 (2009)
33. Fitzgibbon, A., Zisserman, A.: On affine invariant clustering and automatic cast listing in movies. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2352, pp. 289–298. Springer, Heidelberg (2002)
34. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55(1), 119–139 (1997)
35. Fukui, K., Yamaguchi, O.: Face recognition using multi-viewpoint patterns for robot vision. In: International Symposium of Robotics Research, pp. 192–201 (2003)
36. Gao, Y., Wang, T., Li, J., Du, Y., Hu, W., Zhang, Y., Ai, H.: Cast indexing for videos by ncuts and page ranking. In: International Conference on Image and Video Retrieval (CIVR), pp. 441–447 (2007)

37. Garcia, C., Delakis, M.: Convolutional face finder: A neural architecture for fast and robust face detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(11), 1408–1423 (2004)
38. Geng, Y., Shan, C., Hao, P.: Square loss based regularized lda for face recognition using image sets. In: IEEE CVPR Workshop on Biometrics, pp. 99–106 (2009)
39. Goh, R., Liu, L., Liu, X., Chen, T.: The cmu face in action (fia) database. In: IEEE International Workshop on Analysis and Modeling of Faces and Gestures (AMFG), pp. 255–263 (2005)
40. Gorodnichy, D.O.: On importance of nose for face tracking. In: IEEE International Conference on Automatic Face & Gesture Recognition (FG), pp. 181–186 (2002)
41. Gorodnichy, D.O.: Facial recognition in video. In: Kittler, J., Nixon, M.S. (eds.) AVBPA 2003. LNCS, vol. 2688, pp. 505–514. Springer, Heidelberg (2003)
42. Gorodnichy, D.O.: Recognizing faces in video requires approaches different from those developed for face recognition in photographs. In: Workshop on Enhancing Information System Security through Biometrics (2004)
43. Gorodnichy, D.O.: Video-based framework for face recognition in video. In: International Workshop on Face Processing in Video, pp. 330–338 (2005)
44. Grabner, H., Grabner, M., Bischof, H.: Real-time tracking via on-line boosting. In: British Machine Vision Conference (BMVC), pp. 47–56 (2006)
45. Grabner, H., Leistner, C., Bischof, H.: Semi-supervised on-line boosting for robust tracking. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part I. LNCS, vol. 5302, pp. 234–247. Springer, Heidelberg (2008)
46. Gross, R., Shi, J.: The cmu motion of body (mobo) database. Tech. rep., Robotics Institute, Carnegie Mellon University (2001)
47. Hadid, A., Pietikäinen, M.: From still image to video-based face recognition: An experimental analysis. In: IEEE International Conference on Automatic Face & Gesture Recognition (FG), pp. 813–818 (2004)
48. Hadid, A., Pietikäinen, M.: An experimental investigation about the integration of facial dynamics in video-based face recognition. Electronic Letters on Computer Vision and Image Analysis 5(1), 1–13 (2005)
49. Hadid, A., Pietikäinen, M.: Combining appearance and motion for face and gender recognition from videos. Pattern Recognition 42(11), 2818–2827 (2009)
50. Hadid, A., Pietikäinen, M., Li, S.: Learning personal specific facial dynamics for face recognition from videos. In: IEEE International Workshop on Analysis and Modeling of Faces and Gestures (AMFG), pp. 1–15 (2007)
51. Hotelling, H.: Relations between two sets of variates. Biometrika 8, 321–377 (1936)
52. Huang, G., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Rep. 07-49, University of Massachusetts, Amherst (2007)
53. Ikizler, N., Duygulu, P.: Person search made easy. In: Leow, W.-K., Lew, M., Chua, T.-S., Ma, W.-Y., Chaisorn, L., Bakker, E.M. (eds.) CIVR 2005. LNCS, vol. 3568, pp. 578–588. Springer, Heidelberg (2005)
54. Jain, A., Ross, A., Prabhakar, S.: An introduction to biometric recognition. IEEE Transactions on Circuits and Systems for Video Technology 14(1), 4–20 (2004)
55. Kim, M., Kumar, S., Pavlovic, V., Rowley, H.: Face tracking and recognition with visual constraints in real-world videos. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8 (2008)
56. Kim, T., Arandjelović, O., Cipolla, R.: Learning over sets using boosted manifold principal angles (bomva). In: British Machine Vision Conference (BMVC), vol. 2, pp. 779–788 (2005)

57. Kim, T.K., Arandjelović, O., Cipolla, R.: Boosted manifold principal angles for image set-based recognition. Pattern Recognition 40(9), 2475–2484 (2007)
58. Kim, T.K., Kittler, J., Cipolla, R.: Learning discriminative canonical correlations for object recognition with image sets. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3953, pp. 251–262. Springer, Heidelberg (2006)
59. Kim, T.K., Kittler, J., Cipolla, R.: Discriminative learning and recognition of image set classes using canonical correlations. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(6), 1005–1018 (2007)
60. Knight, B., Johnson, A.: The role of movement in face recognition. Visual Cognition 4(3), 265–273 (1997)
61. Kong, S., Heo, J., Abidi, B., Paik, J., Abidi, M.: Recent advances in visual and infrared face recognition - a review. Computer Vision and Image Understanding 97(1), 103–135 (2005)
62. Kozakaya, T., Nakaia, H.: Development of a face recognition system on an image processing lsi chip. In: International Workshop on Face Processing in Video, p. 86 (2004)
63. Krueger, V., Zhou, S.: Exemplar-based face recognition from video. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2353, pp. 732–746. Springer, Heidelberg (2002)
64. Lee, K.C., Ho, J., Yang, M.H., Kriegman, D.: Video-based face recognition using probabilistic appearance manifolds. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 313–320 (2003)
65. Lee, K.C., Ho, J., Yang, M.H., Kriegman, D.: Visual tracking and recognition using prababilistic appearance manifolds. Computer Vision and Image Understanding 99(3), 303–331 (2005)
66. Lee, K.C., Kriegman, D.: Online learning of probabilistic appearance manifolds for video-based recognition and tracking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 1, pp. 852–859 (2005)
67. Li, B., Chellappa, R.: Simultaneous tracking and verification via sequential posterior estimation. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 110–117 (2000)
68. Li, B., Chellappa, R.: Face verification through tracking facial features. Journal of Optical Society of America 18(12), 2969–2981 (2001)
69. Li, B., Chellappa, R.: A generic approach to simultaneous tracking and verification in video. IEEE Transactions on Image Processing 11(5), 530–544 (2002)
70. Li, S.Z., Jain, A.K. (eds.): Handbook of Face Recognition. Springer, New York (2005)
71. Li, S.Z., Zhang, Z.: Floatboost learning and statistical face detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 26(9), 1–12 (2004)
72. Li, Y.: Dynamic face models: Construction and applications. Ph.D. thesis, Queen Mary, University of London (2001)
73. Li, Y., Gong, S., Liddell, H.: Recognising trajectories of facial identities using kernel discriminant analysis. In: British Machine Vision Conference (BMVC), pp. 613–622 (2001)
74. Li, Y., Gong, S., Liddell, H.: Video-based online face recognition using identity surfaces. In: IEEE Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-time Systems, pp. 40–46 (2001)
75. Li, Y., Gong, S., Liddell, H.: Constructing facial identity surfaces for recognition. International Journal of Computer Vision 53(1), 71–92 (2003)
76. Lienhart, R., Kuranov, D., Pisarevsky, V.: Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In: DAGM 25th Pattern Recognition Symposium, Madgeburg, Germany, pp. 297–304 (2003)

77. Liu, C.: A bayesian discriminating features method for face detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(6), 725–740 (2003)
78. Liu, W., Li, Z., Tang, X.: Spatial-temporal embedding for statistical face recognition from video. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 374–388. Springer, Heidelberg (2006)
79. Liu, X., Chen, T.: Video-based face recognition using adaptive hidden markov models. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 340–345 (2003)
80. Martinez, A.M.: Recognizing imprecisely localized, partially occluded and expression variant faces from a single sample per class. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(6), 748–763 (2002)
81. Messer, K., Matas, J., Kittler, J., Lüttin, J., Matitre, G.: Xm2vtsdb: The extended m2vts database. In: International Conference on Audio and Video-Based Person Authentication (AVBPA), pp. 72–77 (1999)
82. Nishiyama, M., Yamaguchi, O., Fukui, K.: Face recognition with the multiple constrained mutual subspace method. In: Kanade, T., Jain, A., Ratha, N.K. (eds.) AVBPA 2005. LNCS, vol. 3546, pp. 71–80. Springer, Heidelberg (2005)
83. O'Toole, A., Roark, D., Abdi, H.: Recognizing moving faces: A psychological and neural synthesis. Trends in Cognitive Sciences 6(6), 261–266 (2002)
84. Ozkan, D., Duygulu, P.: Finding people frequently appearing in news. In: International Conference on Image and Video Retrieval (CIVR), pp. 173–182 (2006)
85. Park, U., Jain, A.K., Ross, A.: Face recognition in video: Adaptive fusion of multiple matchers. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8 (2007)
86. Phillips, P.J., Flynn, P.J., Scruggs, T., Bowyer, K.W., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the face recognition grand challenge. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 947–954 (2005)
87. Ramanan, D., Baker, S., Kakade, S.: Leveraging archival video for building face datasets. In: IEEE International Conference on Computer Vision (ICCV), pp. 1–8 (2007)
88. Ross, D., Lim, J., Lin, R.S., Yang, M.H.: Incremental learning for robust visual tracking. International Journal of Computer Vision 77(1-3), 125–141 (2008)
89. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. Science 290, 2323–2326 (2000)
90. Satoh, S.: Comparative evaluation of face sequence matching for content-based video access. In: IEEE International Conference on Automatic Face & Gesture Recognition (FG), pp. 163–168 (2000)
91. Satoh, S., Nakamura, Y., Kanade, T.: Name-it: Naming and detecting faces in news videos. IEEE Transactions on Multimedia 6(1), 22–35 (1999)
92. Schapire, R.E., Singer, Y.: Improved boosting algorithms using confidence-rated predictions. Maching Learning 37(3), 297–336 (1999)
93. Shakhnarovich, G., Fisher, J.W., Darrel, T.: Face recognition from long-term observations. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2352, pp. 851–868. Springer, Heidelberg (2002)
94. Shih, P., Liu, C.: Face detection using discriminating feature analysis and support vector machine. Pattern Recognition 39(2), 260–276 (2006)
95. Sinha, P., Balas, B., Ostrovsky, Y., Russell, R.: Face recognition by humans: Nineteen results all computer vision researchers should know about. Proceedings of the IEEE 94(11), 1948–1962 (2006)

96. Sivic, J., Everingham, M., Zisserman, A.: Person spotting: Video shot retrieval for face sets. In: International Conference on Image and Video Retrieval (CIVR), pp. 226–236 (2005)
97. Sivic, J., Everingham, M., Zisserman, A.: "who are you?" - learning person specific classifiers from video. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1145–1152 (2009)
98. Stallkamp, J., Ekenel, H.K., Stiefelhagen, R.: Video-based face recognition on real-world data. In: IEEE International Conference on Computer Vision (ICCV), pp. 1–8 (2007)
99. Tan, X., Chen, S., Zhou, Z.H., Zhang, F.: Face recognition from a single image per person: A survey. Pattern Recognition 39(9), 1725–1745 (2006)
100. Tang, X., Li, Z.: Frame synchronization and multi-level subspace analysis for video based face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 902–907 (2004)
101. Tang, X., Li, Z.: Audio-guided video-based face recognition. IEEE Transactions on Circuits and Systems for Video Technology 19(7), 955–964 (2009)
102. Tenenbaum, J.B., Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science 290, 2319–2323 (2000)
103. Turk, M., Pentland, A.P.: Face recognition using eigenfaces. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR (1991)
104. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 511–518 (2001)
105. Viola, P., Jones, M.: Robust real-time face detection. International Journal of Computer Vision 57(2), 137–154 (2004)
106. Wang, R., Shan, S., Chen, X., Gao, W.: Manifold-manifold distance with application to face recognition based on image set. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8 (2008)
107. Waring, C., Liu, X.: Face detection using spectral histograms and svms. IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics 35(3), 467–476 (2005)
108. Wolf, L., Shashua, A.: Learning over sets using kernel principal angles. Journal of Machine Learning Research 4(10), 913–931 (2003)
109. Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y.: Robust face recognition via sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(2), 210–227 (2009)
110. Yamaguchi, O., Fukui, K., Maeda, K.: Face recognition using temporal image sequences. In: IEEE International Conference on Automatic Face & Gesture Recognition (FG), pp. 318–323 (1998)
111. Yang, J., Hauptmann, A., Chen, M.Y.: Finding person x: Correlating names with visual appearances. In: International Conference on Image and Video Retrieval (CIVR), pp. 270–278 (2004)
112. Yang, J., Rong, Y., Hauptmann, A.: Multiple-instance learning for labeling faces in broadcasting news videos. In: ACM International Conference on Multimedia, pp. 31–40 (2005)
113. Yang, M.H., Kriegman, D., Ahuja, N.: Detecting faces in images: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(1), 34–58 (2002)
114. Zhang, Y., Martinez, A.M.: From static to video: Face recognition using a probabilistic approach. In: International Workshop on Face Processing in Video, pp. 78–78 (2004)
115. Zhang, Y., Martinez, A.M.: A weighted probabilistic approach to face recognition from multiple images and video sequences. Image and Vision Computing 24(6), 626–638 (2006)

116. Zhang, Y.F., Xu, C., Lu, H., Huang, Y.M.: Character identification in feature-length films using global face-name matching. IEEE Transactions on Multimedia 11(7), 1276–1288 (2009)
117. Zhao, G., Pietikäinen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence 29(6), 915–928 (2007)
118. Zhao, M., Yagnik, J.: Large scale learning and recognition of faces in web videos. In: IEEE International Conference on Automatic Face & Gesture Recognition (FG), pp. 1–7 (2008)
119. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: A literature survey. ACM Computing Surveys 35(4), 399–458 (2004)
120. Zhou, S., Chellappa, R.: Probabilistic human recognition from video. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2352, pp. 681–697. Springer, Heidelberg (2002)
121. Zhou, S., Chellappa, R., Moghaddam, B.: Visual tracking and recognition using appearance-adaptive models in particle filters. IEEE Transactions on Image Processing 13(11), 1491–1506 (2004)
122. Zhou, S., Krueger, V., Chellappa, R.: Face recognition from video: A condensation approach. In: IEEE International Conference on Automatic Face & Gesture Recognition (FG), pp. 221–226 (2002)
123. Zhou, S., Krueger, V., Chellappa, R.: Probabilistic recognition of human faces from video. Computer Vision and Image Understanding 91(1), 214–245 (2003)