International Journal of Pattern Recognition and Artificial Intelligence

FACE RECOGNITION FROM VIDEO: A REVIEW

JEREMIAH R. BARR, KEVIN W. BOWYER, PATRICK J. FLYNN, SOMA BISWAS

Department of Computer Science & Engineering, University of Notre Dame, 384 Fitzpatrick Hall, Notre Dame, Indiana 46556, United States
jbarr1@nd.edu, kwb@cse.nd.edu, flynn@cse.nd.edu, sbiswas@cse.nd.edu

Driven by key law enforcement and commercial applications, research on face recognition from video sources has intensified in recent years. The ensuing results have demonstrated that videos possess unique properties that allow both humans and automated systems to perform recognition accurately in difficult viewing conditions. However, significant research challenges remain as most video based applications do not allow for controlled recordings.

In this survey, we categorize the research in this area and present a broad and deep review of recently proposed methods for overcoming the difficulties encountered in unconstrained settings. We also draw connections between the ways in which humans and current algorithms recognize faces. An overview of the most popular and difficult publicly available face video databases is provided to complement these discussions. Finally, we cover key research challenges and opportunities that lie ahead for the field as a whole.

Keywords: Face recognition; face detection; face tracking; motion modeling; super-resolution; 3D modeling; manifold modeling; surveillance analysis; video understanding

1. Introduction

Research on face recognition from video has intensified throughout the last decade. This body of work has generally focused on achieving accurate face recognition results in significantly degraded viewing conditions.[1,2] In traditional face image acquisition settings, such as passport agencies or police stations, nuisance variables ranging from head pose to facial expression are controlled. In contrast, video surveillance systems cannot be as obtrusive, so the activities of the recorded individuals and the effects of the environment can vary significantly. Numerous performance evaluation efforts have demonstrated that face recognition algorithms that operate well in controlled environments tend to suffer in surveillance contexts.[1,2,3,4] These issues have motivated the development of face recognition algorithms that draw from the wealth of information provided by videos to compensate for the poor viewing conditions encountered in uncontrolled viewing scenarios.

Specifically, Zhou *et al.*[4] assert that videos afford three useful properties that can aid in recognition:

1. A set of observations - a video sequence contains multiple images of the same face that can potentially show how it appears under different conditions.
2. Temporal dynamics - videos contain temporal information that still images do not possess.
3. 3D information - in an extension to the first property, Zhou *et al.*[4] note that a sequence of video frames can display the same object from a number of different angles, i.e. 2D videos implicitly contain 3D geometric information.

Moreover, neurological evidence suggests that humans exploit these properties by using both the structure of facial features and idiosyncratic facial movements to recognize others.[5] Temporal dynamics play an especially strong role in the recognition of familiar people.

Conversely, the following nuisance factors can arise in unconstrained face recognition applications:

- Pose variation - uncontrolled cameras can record non-ideal face shots from a variety of angles, causing the correspondences between pixel locations and points on the face to differ from image to image.
- Illumination variation - an individual may pass underneath lights with a range of relative positions and intensities throughout the course of one or more videos, so that the surface of the face appears different at different times.
- Expression variation - the appearance of the face changes as the facial expression varies.
- Scale variation - the face will occupy larger or smaller areas in the video frames as it moves towards or away from the camera, and, in the worst case, the spatial resolution of the face can decrease to the point where it becomes unrecognizable. Spatial resolution can also depend on the properties of the camera, such as the depth of field of its lens.
- Motion blur - significant blur can obscure the face if the camera exposure time is set too long or the head moves rapidly.
- Occlusion - objects in the environment can block parts of the face, making the tasks of recognizing the face and distinguishing it from the background more difficult.

These factors may cause the differences in appearance between distinct shots of the same person to be greater than those between two people viewed under similar conditions.[6] Although pose and illumination are traditionally regarded as two of the most challenging nuisance factors, Phillips *et al.*[7] present evidence that some of the other factors listed above have nearly as significant of an impact on face recognition performance in uncontrolled contexts.

These properties and problems are well known throughout the academic, commercial and governmental sectors. The Foto-Fahndung report produced by the German BKA presents an evaluation of three different commercial face recognition systems in a railway station surveillance scenario.[1] Subjects were recorded and rec-

ognized as they descended a stairway or escalator. Pose was partially controlled since the video cameras were placed in front of the stairway. However, illumination conditions changed at night because half of the artificial lights were deactivated. The cameras required wider apertures and longer exposure times to compensate for the degradation in illumination. As a result, the video frames contained more motion blur artifacts. The evaluated systems yielded recognition rates around 60% throughout the day, but night-time performance dropped to 10-20% due to problems with blur and illumination.

Similarly, the 2002 Face Recognition Vendor Test (FRVT) report contains an evaluation that compares face recognition from still images to recognition from videos.[3] In an experiment conducted on a face database comprised by still images of 63 subjects, a number of commercial recognition systems for image based face recognition performed worse when tasked with the identification of faces from videos. The still images from the database contained frontal face views, while the videos displayed speaking subjects with varying expressions. The faces in the videos appeared significantly different from those in the database images, causing a large number of recognition errors.

The Multiple Biometric Grand Challenge featured a problem involving face recognition from videos in which illumination, movement and head pose were not controlled.[2] The video dataset included high resolution (1440x1080) and standard resolution (720x480) sequences with subjects walking toward the camera. Out of four state-of-the-art commercial face recognition algorithms, the best performers on the high and standard resolution videos only reached about 70% and 40% verification rates, respectively. All algorithms performed significantly better on the high resolution videos. Resolution dependent performance differences notwithstanding, off-frontal poses were observed to play a large role in the poor verification performance of all of the systems. In each of these studies, the ability to recognize faces suffered due to variations that intensify the differences in appearance between images of the same individual.

Much of the research on face recognition from video has focused on handling these nuisance factors while taking advantage of the unique characteristics of video data. The studies included in this body of work can be broadly categorized into two groups depending on which video properties they exploit, as shown in Fig. 1. *Set-based approaches* treat videos as unordered collections of images and take advantage of the multitude of observations, whereas *sequence-based approaches* explicitly use temporal information to increase efficiency or enable recognition in poor viewing conditions.

Although set-based approaches do not depend on the ordering of face images, they exploit the quantity and variety of observations to achieve robustness to degraded viewing conditions. These methods differ in terms of whether they fuse information over the observations before or after matching.[8,9] Prior to matching, information can be fused across images at the data or feature levels. Super-resolution techniques operate at these levels to enhance the resolution of the face. Similarly,

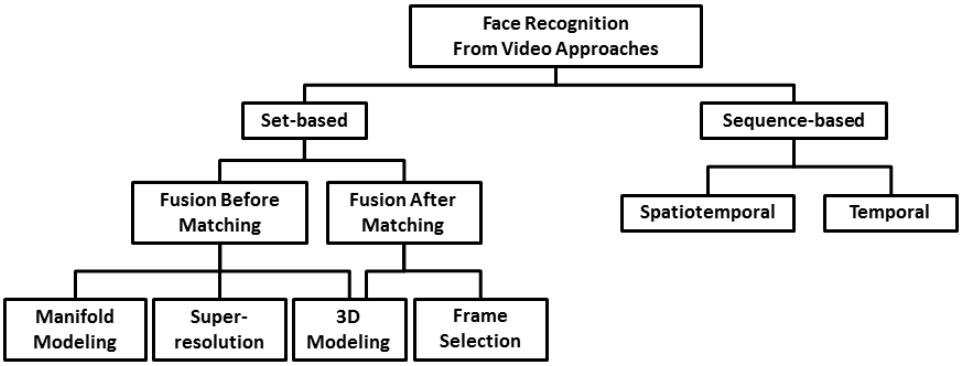

Fig. 1: A Taxonomy of the Face Recognition from Video Literature.

3D modeling techniques leverage the potentially wide range of views contained within a set of frames to recover the 3D structure of face, which can aid in attaining invariance to pose changes. The entire set of observations can be modeled as a manifold or a probability distribution, representations that can potentially allow for robustness to variations in pose, illumination and expression as well as provide the convenience and power of well-established statistical and mathematical techniques. Alternatively, fusion can take place at the level of match scores, ranks or decisions for subsets of images that were deliberately selected to contain a wide variety of appearance conditions or high quality observations.

In contrast to the set-based approaches, sequence-based methods explicitly use temporal cues during recognition. Spatiotemporal techniques leverage both appearance and motion cues to attain a recognition decision, while temporal methods only draw from idiosyncratic facial movements. Sequence-based methods can allow for efficient face tracking and can improve recognition performance in degraded conditions wherein portions of the faces are temporarily deformed, occluded or obscured.[10,11,12,13,14] Whereas set-based methods tend to degrade in performance when presented with facial expression changes, results from Refs. 15, 16 and 17 indicate that a temporally oriented approach can use the ways facial muscles contract and extend when a person speaks or expresses emotions as biometric characteristics.

This survey covers these approaches along with the most popular face video databases and unaddressed research directions. The field of face recognition from video is young relative to that of face recognition from still images, so there is only one prior survey article that is strictly dedicated to the face recognition from video literature.[18] It contains a review of tracking and detection along with a broad overview of the video-based face recognition literature. Zhao *et al.*[6] provide a comprehensive survey on face recognition in general. They decompose the face recognition from video literature into still image, multimodal and spatiotem-

poral categories. In addition, the section on video discusses the face tracking and 3D modeling methods that were the state-of-the-art at the time that survey was written. Zhou *et al.*[4] devote a book chapter to face recognition from video wherein they present a probabilistic identity characterization framework and a literature review encompassing a thorough categorization of the literature based on the properties of videos. Building on these earlier works, this survey includes a broad, deep and up-to-date review of the research on face recognition from video. This review incorporates detailed discussions of results; comparisons between techniques and approaches; notable tie-ins to aspects of human face recognition; and outlines of potential challenges and research directions that lie ahead for the field.

Specifically, Section 2 of this paper discusses publicly available datasets that have been used by a number of research groups. Section 3 covers common biometrics applications and key performance metrics. Next, Section 4 describes the basic face recognition pipeline and the ways in which faces are located and processed prior to recognition. Recent work on set- and sequence-based face recognition from video is covered in Sections 5 and 6, respectively. Finally, Section 7 provides a discussion of open problems and unaddressed research issues.

2. Video Datasets

A number of research groups have amassed datasets with well-defined variations to promote progress in the field. These standard datasets allow for the replication of results along with direct comparisons between methods. The abilities to measure progress and perform systematic research stem from the widespread use of a common basis of comparison. Overviews of some of the most notable datasets are given below.

Goh *et al.*[19] acquired the CMU Face in Action (FIA) database, which contains 640x480 twenty second videos of 214 participants. In the videos, subjects randomly changed their facial expressions and orientations as they enacted a passport enrollment scene. The sequences were acquired at 30 frames per second (FPS) from three different angles in both indoor and outdoor environments. Many of the participants were recorded in three sessions separated by a number of months to allow for time-lapse experiments.

The CMU Motion of Body (MoBo) database acquired by Gross and Shi consists of sequences containing subjects walking on a treadmill.[20] Each of the subjects was recorded with 6 color cameras positioned around the treadmill. Each 640x480 sequence was recorded at 30 frames per second and lasts for 11 seconds. The database includes 150 sequences of 25 subjects either walking slowly, quickly, at an incline or with a ball.

Honda and UCSD acquired two datasets of subjects exhibiting a wide range of poses. The first dataset was acquired by Lee *et al.*[13] and contains 75 videos from 20 subjects. The second dataset was accumulated by Lee *et al.*[14] as well and contains 30 videos from a separate set of 15 subjects. Each 640x480 video was recorded at 15

Table 1: Selected Face Video Datasets. Conditions include: indoor/outdoor (i/o); varying pose (p), illumination (l), expression (e), and scale (s); motion blur (b); occlusions (c); walking (w); random actions and/or motion (r); surveillance quality (v); and multiple people (m).

Dataset Conditions Download Location	Subject Count Resolution	Video Count Frame Rate
CMU FIA[19] i, o, p, l, e,r Contact Goh *et al.*	214 640x480	214 30 FPS
CMU MoBo[20] i, w Contact Gross and Shi	25 640x480	150 30 FPS
First Honda/UCSD[13] i, p `http://vision.ucsd.edu/~leekc/HondaUCSDVideoDatabase/HondaUCSD.html`	20 640x480	75 15 FPS
Second Honda/UCSD[14] i, p `http://vision.ucsd.edu/~leekc/HondaUCSDVideoDatabase/HondaUCSD.html`	15 640x480	30 15 FPS
CamFace[21] i, p, l, r Contact Arandjelović and Cipolla	100 320x240	1400 10 FPS
Faces96[22] i, l, s `http://cswww.essex.ac.uk/mv/allfaces/faces96.html`	152 196x196	152 0.5 FPS
VidTIMIT[23] i, p, e `http://www.itee.uq.edu.au/~conrad/vidtimit/`	43 512x384	43 Frame set
ND-Flip-QO[24] i, o, l, e, s, b, c, r, v, m `http://www.nd.edu/~cvrl/CVRL/Data_Sets.html`	90 640x480	14 30 FPS
YouTube Celebrities[25] i, o, p, l, e, s, b, w, v, m `http://seqam.rutgers.edu/softdata/facedata/facedata.html`	47 180x240, 240x320	1910 25 FPS
MBGC[26] i, o, p, l, e, s, c, w, r, v, m `http://www.nist.gov/itl/iad/ig/mbgc.cfm`	821 720x480,1440x1280	3764 Not described

frames per second and lasts for 15 seconds or longer. Both datasets were recorded indoors which means that illumination variation is not a significant issue.

Table 2: Selected Works Involving Popular Databases. Each of the datasets listed below has been used by multiple research groups to evaluate recognition performance.

Dataset	Approach	Works Presenting Results	Recognition Rate
CMU Face in Action	3D Modeling	Park *et al.*[27]	70%
		Liu and Chen[28]	95.9%
	Frame Selection	Park *et al.*[29]	99%
CMU Motion of Body	Manifold Modeling	Wang *et al.*[30]	96.9%
	Frame Selection	Hadid and Pietikäinen [31]	93.8%
	Spatiotemporal	Hadid and Pietikäinen[32]	97.9%
		Liu and Chen [33]	98.8%
First Honda/UCSD	Manifold Modeling	Zhou and Chellappa [34]	93.3%
		Wang *et al.*[30]	96.9%
		Mian [35]	99.6%
	Frame Selection	Thomas *et al.*[36]	99%
	Spatiotemporal	Aggarwal *et al.*[37]	90%
		Hadid and Pietikäinen[32]	96%

The CamFace database contains 67 male and 33 female subjects of different ages and ethnicities.[21] Every subject has fourteen 320x240 videos that were recorded at 10 frames per second. The clips contain different configurations of multiple light sources. The subjects moved about to create substantial variations in translation, yaw and pitch. However, expression variation is minimal.

Faces96 is comprised by 196x196 image sequences from 152 individuals.[22] The sequences were recorded indoors and contain significant changes in head scale and illumination. All sequences were recorded at 0.5 frames per second during the same day.

The VidTIMIT dataset contains videos of 24 males and 19 females speaking in an office.[23] Subjects rotate their heads in a controlled sequence. A broadcast quality video camera recorded the videos.

Barr *et al.*[24] collected a crowd video dataset containing 14 crowd videos of 90 subjects, five of whom appear in multiple videos and 85 of whom appear in one video. The 640x480 videos were recorded with a Flip camcorder in a variety of indoor and outdoor settings with different illumination characteristics. Subjects were allowed to change their expressions freely.

The YouTube Celebrities dataset, which was collected by Kim *et al.*[25], consists of 1910 noisy YouTube videos of 47 actors and politicians. This dataset is challenging as the majority of the videos are low resolution and highly compressed. Pose, illumination and expression were also largely uncontrolled.

The Multiple Biometric Grand Challenge (MBGC) database includes one of

(a) (b)

(c) (d)

Fig. 2: Selected dataset sequences: (a) MBGC, (b) CMU MoBo, (c) First Honda/UCSD, and (d) YouTube Celebrities.

the largest face video datasets that has been collected to date.[26] This dataset was collected at the University of Texas at Dallas and the University of Notre Dame and spans 3764 visible spectrum videos and 821 subjects. The video set contains a wide variety of videos, some of which show subjects at both frontal and off-angle poses. Other videos display subjects walking, performing activities and conversing with one another in settings with unconstrained illumination, movement and poses. In addition, this dataset contains both indoor and outdoor videos.

The datasets mentioned here vary along a number of important dimensions. The video resolution varies from 196x196 to 1440x1280. Some datasets contain a relatively small number of subjects, *e.g.* the Second Honda/UCSD dataset only contains 15 subjects, while others contain hundreds of individuals, *e.g.* the MBGC dataset has 821 people. Most videos only have a single face in the field of view at any instant, but some datasets contain videos with 10 or more people in view. Similarly, some videos capture the movements of the entire body and show individuals making unpredictable motions, while others focus on the face and display people

performing a limited set of actions. The backgrounds in a portion the datasets are fairly uniform, which simplifies face detection and tracking. Conversely, the backgrounds of other datasets span complex environments with objects that might appear similar to a human face. Comparing the performance results of algorithms operating on different datasets is nearly meaningless due to these large variations along multiple dimensions of difficulty.

3. Applications and Performance Metrics

Research in the area of face recognition from video has primarily focused on its applications in biometrics. Biometrics entails the use of physical or behavioral characteristics to automatically recognize people. Performing face recognition on videos enables unobtrusive identification in uncontrolled environments, which is ideal for surveillance, video-indexing and web content analysis use cases.

Biometric applications generally involve some combination of the identification, verification and watch list tasks.[3] Much of the face recognition from video literature focuses on identification and verification. In each case, the face recognition system is provided a *gallery* or collection of biometric signatures for known individuals. A set of *probes* containing the biometric signatures of unknown individuals is presented for recognition. The system compares probes to the biometric signatures in the gallery to generate match scores. Alternatively, distances or match probabilities may be produced. The identification and verification tasks differ in terms of their objectives.

Identification entails matching the probe set against the gallery. Identification is a common task in the law enforcement domains in which officials must identify suspects using large mugshot databases. Every probe has at least one matching gallery entry called the *correct match*. For each probe, the face recognition system sorts the gallery entries by the strength of their match. A probe is assigned a rank k if the correct match from the gallery has the kth largest match score.[3] The cumulative match characteristic curve, a common way to display identification performance, plots the percent of probes with rank k or higher over a sequence of k values. The rank-one recognition rate tends to be used to summarize overall performance. All results reported here are rank-one recognition rates or accuracies unless otherwise noted.

In the *verification* task, someone claims that he or she is a particular person. The recognition system verifies this assertion by matching the probe against the gallery entry corresponding to the claimed identity. The system accepts the claim if the match score lies above a predetermined operating threshold, otherwise the claim is rejected.[3] A *false accept* occurs when the recognition system decides a false claim is true and a *false reject* occurs when the system decides a true claim is false. Moreover, the false accept rate (FAR) is the the percentage of probes a system falsely accepts even though their claimed identities are incorrect, while the false reject rate (FRR) is the percentage of probes a system falsely rejects despite the

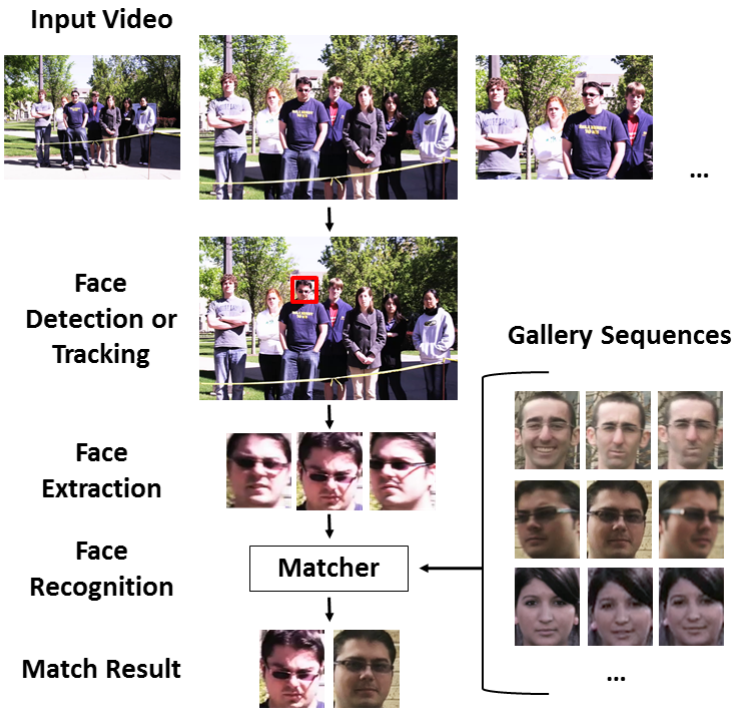

Fig. 3: Basic Recognition Scheme. For each video frame, the faces are first located via a detector or tracker. The located faces are extracted and passed to a matcher, *i.e.* a face recognition algorithm, which measures the similarity between the extracted faces and the faces from a gallery comprised by the sequences of enrolled subjects.

fact that their claimed identities are correct. The performance trade-offs associated with using different parameterizations in a verification system are quantified using both the FRR and the FAR. The receiver operating characteristic (ROC) curve plots the FRR against the FAR as a function of one or more control parameters, including the score threshold. The point where the FRR approximately equals the FAR, termed the equal error rate (EER), is often used to summarize verification performance. A verification algorithm achieves perfect performance if it reaches a 0.0% FRR at a 0.0% FAR.

4. Face Detection, Tracking and Feature Extraction

Face recognition systems typically progress through a number of stages during video processing. The first step is to determine which image regions contain a face via a detection or tracking component. Faces are located by distinguishing facial features

from those of the background. Next, the face information is extracted and converted into the form required by the recognition algorithm, and then the face is matched against the gallery set. Some recognition systems detect or track faces and perform recognition simultaneously.[10,38,11,12,13,14]

Detection or tracking is performed to locate faces if their positions in a video frame are not known prior to analyzing a video. Face detection algorithms treat each image as an independent observation and thus do not model motion state across sequences of video frames. Face tracking algorithms, on the other hand, accumulate both spatial and motion information over subsequences of frames to continuously find image regions containing previously detected faces. In both approaches, an algorithm searches for features in the image that indicate the presence of a face. The difference between detection and tracking lies in the size of the search area.

In Ref. 39, Viola and Jones propose an efficient machine learning approach for combining a small set of features from a large set to detect faces in images. During the training stage, a weighted ensemble of weak classifiers is trained to distinguish faces from other objects, where each weak classifier operates on a particular feature. A variant of the AdaBoost learning algorithm chooses the weighted combination of weak classifiers and, hence, the combination of features that offers the best classification performance on the training set. The features, Haar-like wavelets, can be computed with a small number of operations by using a novel data structure called the integral image. The resulting detector operates on overlapping windows within input images, determining the approximate locations of faces. Viola and Jones received the esteemed Longuet-Higgins Prize at the 2011 IEEE Conference on Computer Vision and Pattern Recognition for this work as it has made a fundamental impact on the computer vision field at large.[40] Consult Refs. 39, 41, 42 and 43 for comprehensive reviews of the work by Viola and Jones as well as other detection algorithms.

Head pose variations and occlusions can change the appearance of the face, making the detection task more difficult. Pose invariance can be achieved by incorporating a set of pose-specific face detectors into an array at the cost of increased computation time, while occlusion can be handled with a part-based detector. Alternatively, faces can be located with a tracker, an algorithm which exploits the temporal continuity inherent in videos to potentially achieve robustness to pose changes and occlusions. Most trackers use a face detector to find faces initially, but then use appearance and motion cues in subsequent video frames. Trackers tend to offer efficiency gains over detectors because they typically do not scan the entirety of every video frame searching for faces.

Deterministic tracking approaches typically optimize some cost function. For instance, the mean shift algorithm can be used to reduce a cost function of the color histogram of a tracked object.[44] Xu and Roy-Chowdhury use a bilinear subspace consisting of illumination and motion variables and an iterative mean-squared error reduction scheme to recover the 3D structure, motion and lighting of an object.[45] In contrast, stochastic tracking techniques, such as the Kalman filter used

by Azarbayejani and Pentland (see Ref. 46) or the more general particle filter employed by Zhou *et al.*[11,12,38], model a hypothesis or a set of hypotheses about the kinematic state of an object. Stochastic algorithms are more robust than deterministic algorithms in the sense that they can avoid local optima of the cost function. Conversely, the deterministic approach tends to be more computationally efficient.

Zhou *et al.*[11,12,38], Li *et al.*[10] and Lee *et al.*[13,14] propagate probability densities over time in schemes that track motion state and perform recognition simultaneously. In particular, Zhou *et al.* and Li *et al.* employ sequential importance sampling techniques to estimate the location and in-plane orientation of faces, while Lee *et al.* use a Bayesian inference framework to recursively fuse the recognition results from each video frame. The associated trackers find the region of the next frame with the shortest distance to the linear subspace representing the face identified in the current frame. Using the same model for both tracking and recognition can result in better face alignment. Moreover, computational costs can be further reduced by representing previously recognized individuals with appearance models.[14]

Tracking and detection algorithms may only provide a coarse estimation of where a face is a located. A more precise boundary can be determined through a process known as localization. A detailed boundary can be obtained with skin detection techniques, which segment the image region containing the face based on pixel color values.[47]

Once located, the image regions containing faces must be extracted and then transformed into the form required for recognition. Typical transformations normalize the face by aligning key facial points, such as the eye centers, to canonical positions; warping the face to compensate for out-of-plane rotation; or smoothing the color or intensity distribution of the pixels in the facial region. Local features, which characterize the information around a set of points, or holistic features, which characterize the appearance of the entire face, are then extracted to form face patterns and passed to the recognition algorithm. These features may also be incorporated into a person-specific model that a tracker can use to locate a particular face in later frames.

At a coarse level, the recognition algorithm can exploit the large number of patterns from a sequence in one of two ways. Set-based algorithms discard the temporal dimension yet take advantage of the multitude of available face patterns. Sequence-based approaches explicitly incorporate temporal information into recognition decisions, with the objective of increasing computational efficiency, improving robustness to nuisance factors, or using facial motion cues as biometric characteristics. Overviews of these approaches are provided in the following sections.

5. Set-Based Approach

The set-based approach poses the face recognition from video problem in terms of matching with sets of multiple samples. Set-based algorithms fuse information over the sample set before or after matching individual face images (see Ref. 8 for an in-

depth discussion of fusion approaches). Information fusion allows a set-based face recognition algorithm to attain higher recognition accuracy, increased robustness to nuisance factors or increased efficiency relative to algorithms for face recognition from still images.

- *Fusion before matching* - the data or features extracted from each face image can be aggregated together prior to recognition. The features or pixel values from an individual image can be unraveled to form a vector; concatenation of the vectors from different frames can yield a single vector with the information from an entire set. A major drawback of this naïve representation stems from its sensitivity to the number of faces and the order in which the vectors are concatenated together.

  In contrast, super-resolution methods attempt to recover high frequency image content from the aggregated frames with the objective of constructing high resolution images. Some 3D modeling techniques also draw data from multiple frames, only with the goal of approximating the geometric structure of the face to achieve pose invariance. In addition, the entire set of faces can be represented with linear subspaces or nonlinear manifolds,[4] constructs with well-defined metrics that measure distances between sets or the variations that they share in common.

- *Fusion after matching* - Pose, illumination and expression variations complicate face recognition by effecting how the face appears. Image sets can be sampled via frame selection algorithms to increase the likelihood that the probe and gallery sets will have similar compositions with respect to the nuisance factors. Additionally, some techniques for achieving pose invariance use 3D head models to synthesize gallery images with similar orientations to the faces in the set of probe images.

  These techniques can be complemented by score, rank or decision level fusion schemes that integrate information over the probe and gallery images to produce a single match decision. In score level fusion, the match scores across the probe frames are combined via summation, multiplication, or by taking the minimum or maximum score for each gallery entry. The estimated identity corresponds to the gallery entry with the highest fused score. The rank level fusion method first ranks gallery entries by their match score in descending order for each frame. The gallery entry with the lowest sum of ranks over the frame set serves as the estimated identity. Finally, decision level fusion is performed by assigning a vote to the gallery entry with the best match score for each face image from the set. The estimated identity is the gallery entry with the most votes.

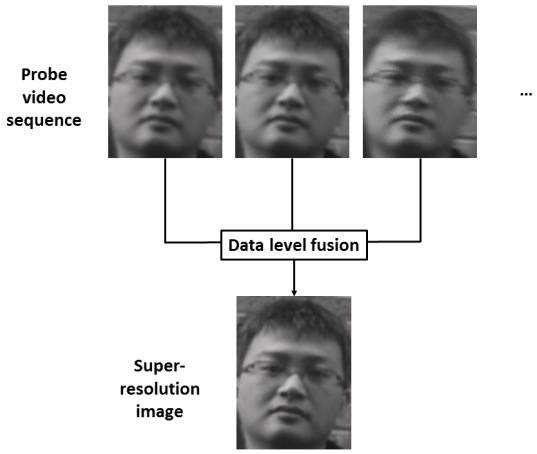

Fig. 4: Super-resolution process. For this example, the spatial information from a set of adjacent video frames was fused together using the iterative back project super-resolution algorithm.

This section covers recent literature on set-based face recognition from video, beginning with the works involving super-resolution and proceeding to those that incorporate 3D modeling, manifold modeling and frame selection algorithms.

5.1. *Super-resolution Methods*

Face recognition performance suffers when the facial resolution drops below the operating range of the recognition algorithm.[48,49] Super-resolution techniques can be applied to recover the high frequency content that was lost due to the limitations of the imaging system.[50] The super-resolution problem can be posed as an inverse problem where the set of observed low resolution images are used to estimate how the captured scene appears. That is, the image formation process can be described with a linear model:

$$y(t) = M(t)x(t) + u(t), \tag{1}$$

where t denotes the time of recording for the observed image sequence; vector $y(t)$ contains the unraveled image content from the observations; the system matrix $M(t)$ captures the effects of motion, sampling and the point spread function of the sensor; vector $x(t)$ represents a sequence of views of the original, high resolution scene; and $u(t)$ models noise. The objective is to obtain an estimate $\hat{x}$ of how the high resolution scene appears. This can be accomplished by minimizing a reconstruction cost function such as the least-squares error,

Table 3: Selected Super-Resolution Approach Results. Conditions include indoor/outdoor (i/o), varying pose (p), illumination (l), and random motions (r). All performance results are given as rank-one recognition rates.

Author, Year Face Representation	Subject Count, Video Count, Resolution Recognition Method	Conditions Performance
Gunturk *et al.*,2003[49] Super-resolution PCA features	68, 68, 40x40[a] Nearest neighbor matching	i,e 74%
Arandjelović and Cipolla,2007[53] Pose and downsampling models	100, 700, 320x240[b] Probabilistic matching	i,p,l,r 95.8%[c]
Al-Azzeh *et al.*,2008[54] PCA features	50, 50, 160x120 Nearest neighbor matching	Not described 97%

Note:
[a]The authors used a gallery set of 68 still images and downsampled the videos to the 40x40 resolution.
[b]The authors used the CamFace dataset.
[c]This performance was attained on 20x20 face images.

$$\hat{x} = \underset{x}{\operatorname{argmin}} \|y - Mx\|_2^2, \tag{2}$$

which uses the L2-norm as measure of the reconstruction error.

Obtaining a reasonable super-resolution estimate is complicated by the fact that this cost function is highly sensitive to noise. Additionally, solving the inverse problem involves inverting the results of applying the system matrix M. Inverting this matrix poses numerical stability issues and, because M is often singular in practical applications, there can be an infinite number of solutions that minimize the cost function. A regularization term can be added to the cost in order to constrain the space of possible solutions and mitigate noise. It is common to use a penalty term that enforces a smoothness constraint, which dictates that the content in x should vary coherently without abrupt changes. This constraint is often incorporated as a *smoothness prior*. Consult Refs. 51, 52 and 50 for comprehensive reviews of super-resolution methods.

Irani and Peleg formalized the iterative back-projection (IBP) algorithm,[55] one of the most popular super-resolution techniques in the field of face recognition from video. The low resolution images must be initially aligned so that salient features occupy approximately the same image locations. The high resolution image is obtained by back projecting the differences between the aligned low resolution observations and low resolution images that are synthesized by applying blur. This back projection process is performed iteratively with the goal of minimizing the

reconstruction cost. Zhou and Bhanu applied the IBP super-resolution technique to recover lost curvature details from low resolution face profile videos.[56] In a similar fashion, Al-Azzeh *et al.*[54] combine this super-resolution scheme with an efficient frequency based alignment procedure that minimizes the warping error between the observations with respect to phase shift and image plane orientation. The super-resolution algorithm drove a principal component analysis (PCA) based matcher to reach a 97% recognition rate in an experiment involving 50 video clips from 50 subjects, which marks an improvement over the 89% recognition rate achieved by the same matcher on native resolution images.

Face recognition systems typically perform dimensionality reduction techniques such as PCA, but in super-resolution schemes like the one employed in Ref. 54, the high resolution images are obtained from the low resolution frames prior to dimensionality reduction. Computational efficiency gains can be achieved by reversing the order of these processes. Gunturk *et al.*[49] transfer the super-resolution problem from the pixel domain to a PCA face space and thus avoid processing image information that will eventually become superfluous. The transformed problem is complicated by two primary noise sources: the image sensor and the representational error incurred by PCA. Face image examples are used to model the statistics of these noise sources. In this work, the reconstruction algorithm uses Bayesian estimation to select the super-resolution PCA feature vector that maximizes the *a posteriori* probability for the set of observed feature vectors from the video frames. Substitution of model-based information into the super-resolution algorithm yields robustness to sensor noise, misalignments and representational error. Tests were conducted using a portion of the CMU Pose, Illumination, and Expression (PIE) database comprised by 68 two second video clips of 68 subjects. The gallery set consisted of neutral expression still images of all of the subjects. The input videos were downsampled to a resolution of 40 x 40 pixels and passed through the super-resolution scheme. The proposed algorithm reached a 74% rank-one recognition rate. For comparison, a traditional PCA based nearest neighbor matcher attained a 79% recognition rate on the original high resolution videos and a 44% rank-one recognition rate on the downsampled videos.

Jillela and Ross select the best frames for a specified super-resolution method instead of imposing model-based constraints.[57] Significant motion blur complicates super-resolution by smoothing corner points that are used for registration. Further, blurry image regions can cause super-resolution algorithms to generate smoothing artifacts. The frame-selection algorithm in Ref. 57 abates these complications by discarding blurry frames. Specifically, an inter-frame motion parameter, β, is derived to measure the average intensity displacement differences between points that are aligned using the Lucas-Kanade optical flow method. If β lies below a given threshold for a pair of consecutive frames, there is no significant motion between these frames and so they are incorporated into the super-resolution process. Otherwise, the frames could introduce artifacts and thus are discarded.

The super-resolution schemes discussed above generally impose a smoothness

prior to regularize the ill-posed nature of the inverse problem. Further, these methods are based on the constraint that the super-resolution images can serve as the bases for reconstructing the low resolution input images when they are transformed according to the image formation model. In Ref. 58, Baker and Kanade provide analytical results showing that the amount of usable information captured by the reconstruction constraints decreases when the magnification factor increases. More significantly, they give empirical evidence that a smoothness prior leads to overly smooth results lacking in high-frequency content at sufficiently large magnification factors, regardless of the number of low resolution observations. They solve these issues by learning models of the relationship between low-resolution face images and their known high-resolution images through a training process. The models are applied to *hallucinate* face images from low-resolution images.

Similarly, Arandjelović and Cipolla propose an approach that does not perform super-resolution explicitly, but instead learns subsampling artifacts on a class-by-class basis.[53] A statistical model of generic face appearance variation is learned offline to characterize the appearance changes due to illumination. The recognition algorithm uses this generic model to re-illuminate the probe sequence and fit it to a probabilistic, person-specific model of downsampling artifacts. In addition to robustness to illumination and downsampling, a robust match likelihood measure provides invariance to changes in head pose. These methods are extended to incorporate a hierarchy of downsampling models that varies over scale, which increases recognition accuracy for arbitrary low-resolution probe images over that of a fixed resolution model. The extended algorithm achieved a recognition rate of 95.8% on videos from the CamFace dataset that were downsampled so that all of the faces had a 20 x 20 pixel resolution.

The model-based approaches proposed in Refs. 58 and 53 counter the information loss due to blur and downsampling without resorting to overly restrictive constraints. They do not, however, address the computational costs associated with aligning images or solving the inverse problem. These costs can make the fruitful application of super-resolution techniques prohibitively time consuming for real-time surveillance applications. Moreover, drawing from cues about the way humans recognize people, psychological studies suggest that high frequency content alone is not sufficient for humans to accurately recognize faces.[59]

5.2. *3D Modeling*

Although the loss of spatial resolution can drastically affect performance, head pose can potentially affect face recognition algorithms more than any other complicating factor.[3] The state-of-the-art commercial systems evaluated during the 2008 Multi-Biometrics Grand Challenge struggled with recognizing faces with off-frontal head poses.[2] The underlying problem is that the correspondence between points on the facial surface and the pixels of its image changes as the head rotates. As a result, a specific pixel location in images of faces with different poses will generally cover

Table 4: Selected 3D Modeling Approach Results. Conditions include indoor/outdoor (i/o), varying pose (p), illumination (l), expression (e), and random motions (r). All performance results are given as rank-one recognition rates unless otherwise noted.

Author, Year Face Representation	Subject Count, Video Count, Resolution Recognition Method	Conditions Performance
Park *et al.*,2007[27] Model recovered via SfM	197, 197, 640x480[a] FaceVACS from Cognitec	i/o,p,l,e 70%
Liu and Chen,2007[28] Face mosaics	29, 290, 640x480[b] Probablistic propagation	i/o,p,l,e 95.9%
Xu *et al.*,2008[60] Projected 3D head model	57, 57, Not described Distance fusion	i,p,l,r 100% max

Note:
[a]The authors used a subset of the CMU FIA dataset containing significant pose variations.
[b]The videos of 29 subjects from the CMU FIA dataset were split into multiple sequences, such that each subject had 10 test sequences.

distinct facial features. In other words, the differences in appearance between images of the same head holding different poses can be greater than those between images of different people with the same head pose.[6]

There are three primary methods for overcoming the pose variation problem: *View synthesis* techniques handle pose differences by rendering face images or sequences with similar poses to the data up for comparison, while *model comparison* techniques match 3D face representations directly. Finally, *view selection* involves acquiring a gallery set with a diverse set of poses for every subject, thereby addressing the problem at the time of enrollment. The former two methods, both of which incorporate 3D face models to varying extents, are discussed in this section, whereas works on the latter approach are covered in Section 5.4.

View synthesis requires a model of the face which is either obtained during gallery enrollment with a 3D sensor or algorithmically synthesized from gallery or probe sequences. Park *et al.*[61] synthesize videos from 3D face models in the gallery to achieve both pose and illumination invariance. A set of SVMs estimates the head pose and illumination conditions present in each of the probe frames, after which the 3D models from the gallery are rotated and lit according to the estimates. All of the models are then projected onto 2D video frames to generate a set of synthetic gallery sequences that match the probe video in terms of the modeled factors. In Ref. 62, Thomas *et al.* present a pose synthesis process whereby generic 3D face models textured with high resolution gallery images are rendered with a variety of pose angles. This method increased the recognition rate of the Viisage

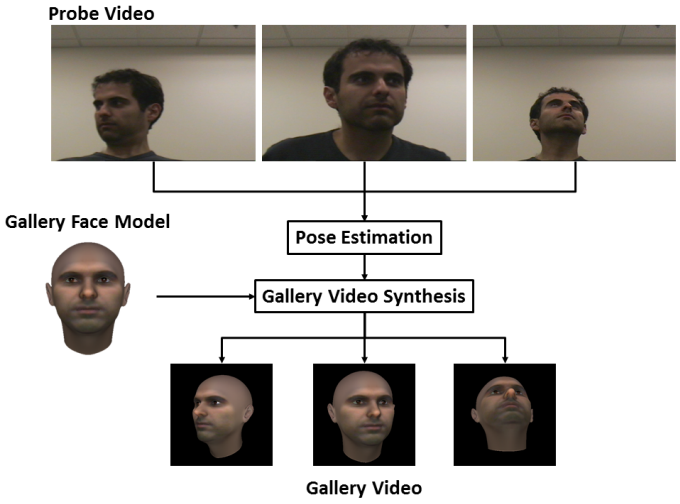

Fig. 5: View synthesis process. In this case, a video synthesis algorithm exploits a combination of pose estimates and a 3D face model from the gallery to synthesize a gallery video with approximately the same head poses as the probe video. The recognition algorithm subsequently matches the probe frames against the corresponding gallery frames. Note: The top row of images was directly extracted from First Honda/UCSD videos.

IdentityEXPLORER matcher on a challenging dataset with surveillance videos of 57 subjects.

Similarly, Xu *et al.*[60] employ a bilinear model to synthesize gallery sequences with the same poses and lighting conditions as probe videos using 3D models that are constructed algorithmically. The bilinear model incorporates facial motion and structure information along with a set of nine basis images defined using spherical harmonics. An inverse composition algorithm estimates the illumination and motion parameters, which in turn are used to synthesize video sequences with 3D face models from the gallery. This framework was evaluated on a private dataset containing 57 subjects rotating their heads under varied lighting conditions. A 100% rank one recognition rate was obtained when the average pose angle was 15 degrees from frontal. The rank one recognition rate degraded to 93% when the average pose angle increased to 45 degrees from frontal.

Storing 3D models in the gallery offers the advantage that range scanners or algorithms can be used to capture or generate 3D data offline, regardless of whether the probe data will be acquired in uncooperative contexts. However, the synthesis process is computationally burdensome as it entails rendering an image of every gallery model for each probe video frame, making it difficult to scale this approach to datasets spanning thousands of individuals. A more efficient way to perform view

synthesis is to recover the 3D structure of the face from probe sequences and render the probe faces to match the gallery data in terms of pose and other factors.

Structure from motion (SfM) is a data level fusion technique that exploits the multitude of views of the same object to recover its shape. SfM algorithms first associate corresponding image points from different views and then recover their 3D point coordinates by solving the inverse problem of how the observed image was formed.[63] Point tracking techniques solve the correspondence problem by analyzing the history of point motions.

Park and Jain apply SfM as a means of recovering 3D models from probe face sequences.[27] A generic active appearance model (AAM), which serves as the basis of a point tracker, is trained on a variety of face views. The AAM points that were localized in one frame serve as seed points for the next frame. The factorization method is applied to recover the face shape from the AAM under the assumptions that the human face is a rigid object and any changes due to facial expression constitute noise. Although the factorization method yields a pose estimate for each point in the appearance model, the estimate lacks a third coordinate. Hence, pose estimates are attained through a gradient descent algorithm that iteratively fits the constructed 3D face shape to the 2D feature points from the AAM. A subsequent texture mapping produces a 3D face model that is then rotated to frontal view and projected onto a 2D image plane. The recognition scheme incorporates the FaceVACS matcher from Cognitec, which is used to compare the synthesized probe images to frontal view gallery images. Experiments were performed on videos from the CMU FIA dataset. Although the 3D model reconstruction algorithm failed on 24 out of the 221 subjects from the data set, the recognition rate on the frontal pose frames of the remaining individuals was nearly 100%. When using frames containing non-frontal faces, the proposed 3D modeling method increased the recognition rate of the FaceVACS matcher from 30% to about 70%.

Although this result suggests that extracting 3D models from the probe images can partially mitigate the pose problem, it also indicates that the 3D shape recovery problem is still open in the domain of video-based face analysis. In Ref. 4, Zhou and Chellappa present three complications that arise from the application of SfM methods to 3D face modeling:

1. the perspective camera model is not well-posed;
2. the shape of the face is not constant due to facial expressions, but SfM works best on rigid objects;
3. the SfM algorithm usually begins with a sparse set of points provided by a point tracking algorithm, and thus must use interpolation to produce a dense set from points that were potentially tracked poorly.

These challenges have played a role in the development of alternative methods to recover the geometric attributes of the face that do not resort to synthesis-driven image matching. Instead, these techniques model probe and gallery sequences alike and avoid the complications associated with SfM via disparate geometric repre-

sentations. For example, Krüger *et al.*[64] combine Bayesian probability propagation with appearance-based 3D models to perform tracking and recognition. The 3D models are constructed with eigen light-fields, a 3D representation that allows the recognition and tracking algorithm to handle out-of-plane rotation along with affine transformations in the image plane.

Liu and Chen model facial appearance and geometry with face mosaics.[28] A 3D ellipsoid upon which face images are back projected approximates the head; a texture map is formed subsequent to back projection. The texture map is decomposed into patches that can move locally so that the same feature points in different images, *e.g.* the corners of the mouth, can occupy the same patch location. Corresponding patches from the training images of a subject are used to learn an array of patch-specific PCA models that characterize face appearance. In addition, a PCA model of the deviations learned from patch movements characterizes face shape. The residues between corresponding patches in a testing and training model pair are fed into a probabilistic distance model. The similarity score is given by the average of the probabilities over all patches. Finally, the CONDENSATION-based framework proposed by Zhou *et al.*[38] is used to combine the face tracking and recognition processes. This face mosaicing method achieved a 95.9% recognition rate on a 29-subject subset of the CMU FIA database.

The 3D modeling approaches discussed above compensate for pose and, in some cases, illumination variations under the assumptions that the pose and illumination estimations and the image registration processes that occur after synthesis are accurate. As Chellappa *et al.* discuss in Ref. 65, accurate registration is critical to handling pose variation. In one instance, the rank-one recognition rate of the view synthesis approach proposed by Xu *et al.*[60] degraded from 100% to 93% when the average pose angle increased from 0 to 45 degrees as a result of inaccuracies in the pose and illumination estimates and registration errors. Additional complications can arise for modeling methods as approximations must be made while forming the models. In particular, strategies that model both probe and gallery sequences necessitate two approximations, one for the probe model and one for the gallery model. This scheme allows errors to occur on both sides of the recognition task.

Finally, the density of the point sets associated with shape estimation decreases when high frequency image content isn't available. Low resolution videos can reduce the effectiveness of model based techniques to the point where modeling is not possible. The model based methods proposed by Lie and Chen in Ref. 28 and Krüger *et al.* in Ref. 64 increase the likelihood of attaining a result by incorporating spatiotemporal techniques, which integrate evidence over time. Arandjelović and Cipolla model head pose and the image sampling process alike in order to compensate for a lack of high frequency content.[53]

5.3. *Manifold Modeling*

The super-resolution and 3D model based methods exploit the multitude of observations afforded by videos through direct means of modeling the image formation process. In contrast, the manifold methods fuse the image data to characterize the space of faces without directly accounting for the image formation process. Manifold models compactly characterize the relationships between faces in general.

The face manifold is comprised by the collection of possible face images or feature patterns within some space.[66] The appearance of the face is a function of its configuration, *i.e.* its pose, expression, scale and so forth. The face manifold X_i can be expressed as

$$X_i = \{c(x_i) : c \in C\}, \tag{3}$$

where C denotes the set of possible face configurations and $c(x_i)$ represents a face image of individual i with configuration c. The union of all person-specific face manifolds forms the manifold of all faces, X. In general, both X_i and X are nonlinear; have a lower dimensionality than the space containing the input data; and can be approximated by a collection of linear subspaces.[30]

Early attempts to characterize the face manifold used PCA to compactly represent the set of faces as a linear subspace. In Ref. 67, Yamaguchi *et al.* present the mutual subspace method (MSM). The MSM forms linear subspaces from entire face sequences via PCA. The associated similarity metric measures the smallest principal angle between two subspaces, *i.e.* the minimum angle between the principal component vectors in each subspace. The cosine of this angle indicates the similarity of the primary mode of variation that is shared by both subspaces.[30] Experiments were conducted on a private 101 person database with training and testing subsets. MSM outperformed a single image based approach by reaching an equal error rate of about 5%.

Arandjelović and Cipolla apply the MSM in a framework that weights the match score contributions of subspaces built from intensity features and subspaces built from illumination invariant feature representations, such as self-quotient images.[71] First, the similarity score between a pair of subspaces built from two intensity video sequences is computed and normalized. The similarity score for the corresponding subspaces from a specified quasi-illumination invariant feature space is computed and normalized next. The final similarity measure is given by the weighted average of these scores. The weighting scheme provides robustness to lighting and shadow variations by using quasi-illumination invariant features when the illumination conditions of a face sequence pair differ drastically. The features derived from intensity images receive more weight when the illumination distributions are similar, as the quasi-illumination invariant representations can introduce errors and artifacts. On a 60 person subset of a large video database comprised by 323 subjects and 1474 videos spanning the CamFace, Faces96 and other databases, a recognition rate of

Table 5: Selected Manifold Modeling Approach Results. Conditions include indoor/outdoor (i/o); varying pose (p), illumination (l), and expression (e); walking (w); and random motions (r). All performance results are given as rank-one recognition rates unless otherwise noted.

Author, Year Face Representation	Subject Count, Video Count, Resolution Recognition Method	Conditions Performance
Yamaguchi *et al.*,1998[67] PCA subspace	101, 101, 180x101 faces MSM	i,p,e 5% EER
Shakhnarovich *et al.*,2002[68] Normal distributions	29, 29, 22x22 faces Symmetric KL divergence	p 100%
Fan and Yeung,2006[69] Isomap affinity matrix	40, 80, Not Described HAC to obtain personal subspaces	p,l,e 95.6%
Zhou and Chellappa,2006[34] Probability distributions	20, 75, 640x480[a] PDF distance measures	i,p 93.3%
Arandjelović and Cipolla,2006[70] Normal distributions	100, 200, Not described Resistor-average distance	i,p,r 98%
Wang *et al.*,2008[30] Appearance manifolds - -	20, 75, 640x480[a] Manifold-manifold distance 25, 150, 640x480[b] -	i,p 96.9% i,w 93.6%
Arandjelović and Cipolla,2009[71] Quasi-illumination invariant features	323, 1474, 160x120-320x240[c] Constrained MSM	p,l,e,r 97%[d]
Lina *et al.*,2009[72] View-dependent covariance matrices	20, 60, Not described Mahalanobis distance	p,e 98% max
Hadid and Pietikäinen,2009[66] LLE manifold	43, 43, 512x384[e] Manifold distance	i,p,e 99.8%

Note:
[a]Experiments were performed on the First Honda/UCSD dataset.
[b]The authors performed experiments on the CMU MoBo dataset.
[c]This dataset was comprised by the CamFace, Faces96 and other databases.
[d]This result was obtained on a 60 subject subset of the dataset.
[e]The authors used the VidTIMIT database.

97% was achieved with constrained MSM (CMSM) and a self-quotient image filter. This result indicates that the MSM provides flexibility with respect to the representation of the input data. Moreover, the associated similarity functions are simple and computationally efficient.

Methods that use a single linear subspace to characterize the observations from

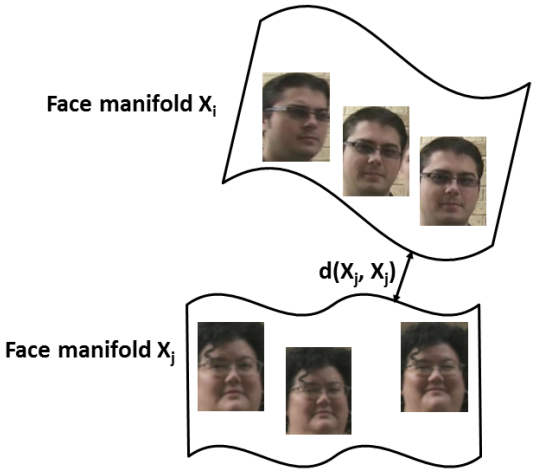

Fig. 6: Manifold to Manifold Distance Computation. In this example, the distance d between manifolds X_i and X_j of identities i and j is measured as the minimal distance between points in pairs spanning both manifolds.

an entire sequence, however, cannot accurately account for nonlinear appearance variations, such as those caused by multiple light sources.[66] The nonlinear structure of the face manifold can be approximated with a set of linear subspaces as the local linearity property holds everywhere on a globally nonlinear manifold.[73] Wang *et al.*[30] model the manifold with a collection of local linear subspaces called maximal linear patches (MLPs) and compute manifold distances by integrating over the distances between MLPs. Rather than clustering a particular set of images to form the subsets from which to construct the local linear subspaces, which requires that the number of clusters be specified *a priori*, the authors propose a one-shot clustering method that incrementally constructs each new MLP from a seed point until the linearity constraint is broken. In this way, the manifold modeling algorithm determines the number of local models adaptively. The distance metric for a pair of MLPs measures dissimilarity in terms of modes of variation, as quantified by the mutual subspace angle, and the distance between constituent data points, as approximated by the distance between exemplars. The manifold-manifold comparison method reached 96.9% recognition accuracy on a subset of 59 videos from the first Honda/UCSD dataset and 93.6% recognition accuracy on the CMU MoBo dataset.

Lina *et al.*[72] propose a similar modeling method. Their scheme embeds view-dependent covariance matrices in each manifold X_i and applies interpolation to approximate unseen poses. The embedding process begins by building a global eigenspace with PCA to reduce the dimensionality of the facial features from all individuals and for all pose angles. Then, for each training pose and individual, a

mean vector and covariance matrix is calculated. Synthetic transformations, including geometric warping and other noise producing procedures, yield a large enough sample set to obtain reliable estimates of these statistics. An additional manifold interpolation step facilitates the calculation of the covariance matrices and mean vectors for poses that lie outside the range of the training data. Manifold matching is performed by measuring the Mahalanobis distances between gallery manifold sections and probe images with like poses. A supervised version of the manifold learning algorithm reached 98% accuracy when trained and tested on different subsets of a small, private database with 60 videos and 20 subjects.

Fan and Yeung perform hierarchical agglomerative clustering using the geodesic distances between face points lying on the manifold.[69] Each cluster is represented by an intrapersonal and an extrapersonal subspace. The intrapersonal subspace expresses the variation within a cluster, whereas the extrapersonal subspace expresses the variation between a cluster's images and the exemplars of nearby clusters. The distance between a probe face image and a gallery cluster is measured as the angle between the projections of the image onto the intrapersonal and extrapersonal subspaces. A 95.6% average recognition rate was obtained on the evaluation dataset used in their earlier work (see the discussion on Ref. 74 in Section 5.4 for details on the dataset).

Hadid and Pietikäinen combine manifold learning with a novel manifold distance measure to match sequences against sequences.[66] Locally linear embedding (LLE) is executed independently on each subject-specific training video, with the objective of recovering the low-dimensional face manifold X_i from the high-dimensional face image space. A test sequence is projected into the embedding space of each training manifold during recognition, which yields a collection of test manifolds. The training manifold that lies closest to its corresponding test manifold represents the best match. Experiments conducted on the VidTIMIT dataset compared this approach to still-image recognition methods based on PCA, LDA and local binary patterns (LBP) as well as two spatiotemporal techniques that incorporated hidden Markov models (HMM) and auto-regressive and moving average (ARMA) models. Face images were deliberately extracted from each video using the eye positions detected in the first frame so that the faces were poorly aligned. The proposed method handled this complication successfully and achieved a 99.8% recognition rate. For comparison, the PCA, LDA, LBP, HMM and ARMA techniques attained respective recognition rates of 94.2%, 94.0%, 97.6%, 92.9% and 95.8%.

The manifold representation can be exploited to attain facial appearance probability distributions. Practically speaking, the anatomical structure of the neck limits head motion, which makes some head poses more likely than others. The images from a sequence can thus be treated as independent samples drawn from some face appearance distribution. The distribution representation allows for a probabilistic treatment of noise and outliers, while enabling the use of probability density function (PDF) distance metrics.[4] Specifically, this representation treats a face image x for subject i as a sample drawn from the probability density $p_F^i(x)$.[70] Let f_i denote

a mapping function which embeds faces in the image space and assume that noise drawn from a distribution p_n perturbs the image formation process. The probability of observing an image X can thus be expressed as:

$$p^i(X) = \int p_F^i(x) * p_n(f_i(x) - X)dx. \tag{4}$$

Shakhnarovich *et al.*[68] employ a multivariate normal density to represent face appearances and the Kullback-Leibler (KL) divergence to measure density similarity. In this way, the Kullback-Leibler (KL) divergence is used to express the overall distance between face manifolds drawn from different image sets. The KL divergence measures how well a PDF $q(x)$ accounts for information in the set represented by another PDF, $p(x)$. It is given by:

$$D_{KL}(p||q) = \int p(x) lg \ \frac{p(x)}{q(x)} dx. \tag{5}$$

The performance of the proposed approach was evaluated on a small, private 29-subject dataset containing a single video for each individual. One portion of each sequence was used for training and the other was used for testing. Perfect recognition accuracy was achieved, albeit on strongly correlated training and test sets.

One issue with modeling appearance with a Gaussian distribution is that the model is easily violated by variations due to pose and illumination. Much like Shakhnarovich *et al.*[68], Arandjelović and Cipolla represent face video sequences as sets of independently and identically distributed samples drawn from a PDF and use the KL divergence as a distance metric.[70] They employ the KL divergence as the basis of the symmetric Resistor-Average distance (RAD) metric given below.

$$D_{RAD}(p, q) = (D_{KL}(p||q)^{-1} + D_{KL}(q||p)^{-1})^{-1} \tag{6}$$

No closed form expression for the D_{KL} exists for the general case of estimating the distance between two nonlinear face manifolds, but an analytical expression is possible for normal distributions. Hence, a nonlinear projection of face data via kernel principal component analysis (Kernel PCA) is performed to guarantee the normality of $q(x)$ and $p(x)$. This operation unfolds the face manifolds in the embedding image space, making efficient computation of D_{KL} possible and more numerically stable. Robustness to outliers is achieved via the RANSAC algorithm. The proposed approach attained a 98% recognition rate on a dataset containing 100 subjects and 200 videos, while an MSM based algorithm only reached 89%.

In a similar framework that represents face image sets as sample collections drawn from a probability distribution, Zhou and Chellappa apply a kernel function that computes the inner product of vector pairs that are mapped from the face

space into a high-dimensional feature space via a nonlinear mapping.[34] The kernel function thereby preserves the nonlinearity inherent to the face data. Explicit computation of the mapping is avoided by applying the well-known kernel trick. Further, the nonlinear model is exploited to analytically derive probabilistic distance measures such as the KL divergence that account for higher order statistical properties of the samples. A combination of these techniques and the bag of pixels image representation achieved a 93.3% recognition rate on the First Honda/UCSD data set.

Turaga *et al.*[75] present distance metrics for the Grassmann and Stiefel manifolds, upon which video based face data naturally lies. Additionally, Turaga *et al.* discuss parametric and non-parametric manifold density functions that characterize their geometric structure. They demonstrate that the Procustes distance metric provides computational efficiency, and that non-parametric kernel methods and the Matrix Langevin and Matrix Bingham distributions can be used to approximate class conditional probability densities.

The manifold and probability density representations can lead to recognition failure if the collection of observations is not a sufficient sampling. Hence, formulating the face recognition problem in terms of manifolds or PDFs can preclude the use of short video sequences captured in a wide variety of viewing conditions. Probability density based representations also incur computational complications: The associated distance measures on general distributions may not be numerically stable and empirical methods for characterizing them are computationally intensive.[70]

5.4. *Frame Selection*

The data and feature level fusion methods presented in the prior sections can incur steep computational costs as they generally include all observations regardless of the amount of information they contribute to the recognition result. Another way to exploit the observations available in videos is to select a set of frames that should yield the best recognition accuracy with a given classifier. Frame selection approaches can be used to filter out low quality face images or images that were captured under different conditions than the gallery set. Alternatively, diversity oriented selection techniques can be used to build face image sets that span a wide variety of conditions, which increases the likelihood that the selected probe images will be similar to the gallery images with respect to the nuisance variables.

Quality based approaches generally use statistical approaches to find outliers or draw from heuristics about the pose, lighting, resolution and other characteristics of a useful face image. For example, Berrani and Garcia apply the robust PCA (RobPCA) algorithm proposed by Hubert *et al.*[78] to detect outliers in video sequences that could cause recognition errors.[79] Potential outliers include face images with disruptive illumination effects, off-frontal head poses, poor alignment or any other property that causes them to deviate from a PCA based face model. Such appearance changing influences can force the traditional eigenfaces approach

Table 6: Selected Frame Selection Approach Results. Conditions include indoor/outdoor (i/o); varying pose (p), illumination (l), expression (e), and scale (s); occlusions (c); walking (w); random motions (r); and surveillance quality (v). All performance results are given as rank-one recognition rates unless otherwise noted.

Author, Year Face Representation	Subject Count, Video Count, Resolution Recognition Method	Conditions Performance
Hadid and Pietikäinen,2004[31] Exemplars from LLE space	24, 96, 640x480[a] Nearest neighbor matching	l,r 93.8%
Fan and Yeung,2006[74] Isomap affinity matrix	40, 80, Not Described HAC	p,l,e 96.5%
Zhang and Martinez,2006[76] Region-based PCA,ICA,LDA features	50, 100, Not described[b] Probabilistic fusion over regions	i,p,e,c 99% max
Park *et al.*,2007[29] Face images and PCA features	221, Not described, 640x480[c] PCA and cross-correlation matchers	i,p,e 99% max
Thomas *et al.*,2007[36] Not described	20, 75, 640x480[d] FaceIt SDK	i,p 99%
Stallkamp *et al.*,2007[77] DCT coefficients	41, 2292, Not Described kNN or GMM	i,e,l,s,c,w,r,v 92.5% max
Mian,2008[35] SIFT descriptors	20, 75, 640x480[d] Hierarchical clustering	i,p 99.5%

Note:
[a]The authors performed expirements on the CMU MoBo dataset.
[b]The training set contained three neutral expression still images for each subject. The test set contained two clips for each subject, where the first clip showed the subject randomly talking from a nearly frontal perspective and the second clip captured the subject rotating his or her head.
[c]A subset of indoor videos with restricted illumination variations from the CMU FIA dataset was used for experiments.
[d]The authors used the First Honda/UCSD dataset.

to fit the most informative dimensions to variations caused by noise. The RobPCA algorithm provides robustness to outliers by finding groups of training patterns that occupy compact spaces and by maximizing a robust measure of spread while selecting eigenbases. The proposed algorithm matches the faces that remain after it discards patterns that lie far from the face space.

While the frame selection technique discussed above discards poor quality face images, weighted selection schemes quantify the contribution of each face in a set to the recognition score and assign poor quality face images lower weights. This strategy allows for the inclusion of more information into recognition scores and

decisions. Zhang and Martinez propose a framework that achieves robustness to localization errors, partial occlusion, expression changes and pose variations using a combination of weighting techniques.[76] For each training image, the feature space subset that represents all images under all possible localization errors is estimated with a mixture of Gaussians. Faces are divided into subregions characterized by linear subspaces and Gaussian mixtures, which gives robustness to occlusions and localization errors. The probability of a probe face image matching a gallery class is expressed in terms of the class conditional probabilities of all corresponding subregions in a pair of images; the contributions of the subregions and the entire face are weighed. Experiments were conducted on a private 50 subject database comprised by a gallery set with three neutral expression images for each subject and two 40 frame probe videos of every individual. When using LDA to obtain the subregion feature spaces, the proposed approach attained a maximum recognition rate of 99%.

Stallkamp *et al.*[77] weight the contribution of probe video frames to match probabilities using three techniques. The distance-to-model (DTM) method reduces the influence of probe frames that are significantly different from the closest matching gallery image, *i.e.* the frames that are most likely to cause a misidentification. The second method, distance-to-second closest (DT2ND), addresses situations in which a probe frame could potentially correspond to multiple identities. The third weighting scheme fuses the weights obtained by DT2ND and DTM with the product rule. The weighted frame scores from a particular video are combined using score sum fusion in all weighting schemes. The DTM, DT2ND and combined weighting approaches reached 92.0%, 91.3% and 92.5% respective recognition rates using a kNN matcher on a surveillance database.

Conversely, temporal continuity causes adjacent video frames to contain highly redundant information. Selecting all of the frames in an interval decreases execution efficiency as more comparisons must be made. Consecutive frames also may not bolster recognition improvements because they contain highly correlated information. Some recognition schemes thus incorporate diversity oriented selection algorithms that enforce the construction of sample sets that span as many variations as possible. This approach can potentially yield a compact characterization that includes the significant intra-class variations. Diversity can be achieved with clustering methods, during acquisition or by choosing images with a wide range of qualities according to some metric.

Clustering algorithms can partition a set of images into a collection of groups, where each group contains images with a particular mode of variation.[31] The exemplars that best represent a particular set of appearance variations can subsequently be selected from each cluster as representatives of how someone appears. The mean feature vector for a cluster or a feature vector located near the cluster center can serve as an exemplar. Exemplar-based recognition is popular because it can drastically reduce the number of face comparisons, yet it preservers a significant amount of identifying information.

Fan and Yeung locate the most suitable exemplars from a sample of global face features with an unsupervised non-parametric technique.[74] They employ the Isomap nonlinear dimensionality reduction algorithm to produce a face point affinity matrix for later hierarchical agglomerative clustering. Isomap accounts for the relationships between face images by computing the geodesic distances separating them along the face manifold.[80] The cluster centers determined after hierarchical agglomerative clustering are treated as exemplars, while the recognition decision is made using majority vote fusion to combine matching results. Performance was measured on a small, private 40-subject video database containing variations in pose and illumination. These methods achieved a maximum recognition rate of 96.5%.

Hadid and Pietikäinen employ the Locally Linear Embedding (LLE) algorithm to construct global feature representations of reduced dimensionality.[31] Given a set of face appearance vectors, the unsupervised LLE algorithm maps the vectors onto a neighbor-preserving embedding space of lower dimensionality. K-means clustering is subsequently applied in the embedding space. The centers of the obtained clusters are used to characterize the intra-class variations due to changes in pose, illumination and expression. Matching consists of comparing cropped face images from probe sequences to exemplars from the gallery and performing probabilistic voting to fuse the decisions from multiple frames. On the CMU MoBo database, the combination of LLE with k-means clustering achieved marginally better performance than that offered by a self-organizing map or k-means clustering in an Isomap embedding space. The proposed approach reached a 93.8% recognition rate.

An algorithm proposed by Mian chooses a representative set of local SIFT features from multiple face images with a hierarchical clustering technique and a voting scheme.[35] The SIFT features enable robustness to occlusion and rotation. Face pair similarity is measured in terms of the angle between SIFT vectors and the number of matching SIFT vectors. A weighted average of these measures provides the similarity scores on which the hierarchical clustering algorithm operates. The author chooses a particular partitioning by specifying the number of clusters to form. Each of the clusters contains faces with related appearances, i.e. similar expressions and poses, so that multiple clusters correspond to the same face. A voting process performed during face matching is used to select a representative set of features. A maximum recognition rate of 99.6% was achieved on the First UCSD/Honda dataset with this scheme.

The diversity constraint can also be addressed at enrollment time. Park *et al.*[29] focus on pose variations in a view selection approach by deliberately composing a gallery with a large range of head poses. They also account for motion blur estimates via an analysis of the high frequency components of the discrete cosine transformation of probe video frames. This information enables a view synthesis algorithm to generate gallery videos that match the pose and blur conditions observed in probe sequences. Additional robustness to matcher specific errors is achieved by fusing the results from three matchers: the FaceVACS face recognition application from Cognitec as well as PCA and cross-correlation based matchers produce the

frame set match scores.[81] The individual frame scores are then fused at the score level to attain a single score over all matchers and video frames. Experiments were conducted on a subset of the CMU Face-In-Action (FIA) database that contained significant pose variation, but little illumination variation, as the primary focus was to address pose and blur. The system reached a 99% rank-one recognition rate.

Recognition schemes can employ algorithms for selecting diverse sets of high quality images from video sequences. Thomas *et al.*[36] propose strategies for selecting diverse and high quality image sets in a principled fashion. The first strategy, N highest faceness (NHF), chooses the highest quality frames from a face image sequence based on a quality measure called "faceness" produced by the L-1 Identity Solutions FaceIt face recognition software. The N evenly spaced from M highest faceness (NEHF) strategy sorts the images from a sequence by "faceness" and selects M evenly spaced faces. The last two strategies, largest average distance (LAD) and LAD highest faceness (LADHF), explicitly make diverse selections by choosing faces separated by the largest distances within a PCA feature space. LADHF adds a step wherein face sequence images are ordered by quality prior to the diversity oriented selection. A 99.0% recognition rate was reached on the First Honda/UCSD dataset with the NHF approach.

Xiong and Jaynes present procedures for selecting high-quality mug shot images that are suitable for simple, still image based recognition techniques.[82] They focus on surveillance videos containing challenging variations in pose, illumination and expression. As a result, the proposed system accepts still face images provided that they meet certain intrinsic and extrinsic quality constraints. The intrinsic quality of a face image is given by a function of its orientation, aspect ratio and resolution. The extrinsic quality measure depends on clustering and rewards new face images that increase the density of some mode in the data or shift the mean of the convex hull surrounding a mode away from other classes. The normalized, weighted sum of these measures feeds the decision process that determines whether to add face images to the database. Experiments reported on 96 hours worth of surveillance footage demonstrate the ability of the proposed selection algorithm to support nearest-neighbor based classifier in achieving high accuracy on a challenging data set.

Quality oriented frame selection algorithms can suffer when the assumption that high quality frames exist is violated, which may happen often in surveillance videos and movies for sensitive recognition algorithms. Unlike 3D modeling techniques, these methods cannot generalize over pose and illumination variations that are not present in the gallery or training set, as using a subset of frames equates to drawing samples from a small region of the face space and ignoring potentially useful data from other regions.

6. Sequence-Based Approach

While set-based approaches exploit the potentially large samplings that videos can contain, sequence-based approaches also incorporate the information about their ordering during recognition. Temporal dynamics can be exploited to characterize how facial appearance and motions vary together; improve registration accuracy through a unified tracking and recognition scheme; or represent idiosyncratic features of a person. Unlike in the context of face recognition from still images, video streams can enable a sequence-based algorithm to correctly recognize individuals in contexts that do not give strong support for a decision, such as when the face is intermittently blocked from view.

Temporal continuity plays an important role in human facial and object recognition, as shown by recent psychophysical and neurological studies.[5,83,84] Humans incorporate both static and behavioral facial information during recognition, and, while static information tends to play a stronger role, dynamics aid recognition under poor viewing conditions. In contexts involving the recognition of familiar faces, humans appear to draw heavily from temporal cues when image quality degrades. In a study conducted by Knappmeyer *et al.*[85], participants initially learned to discriminate between two synthetic face models animated with distinct idiosyncratic facial movements before being presented with intermediate morphs between the heads. The identity decisions for the morphed heads were biased by their facial movements. In addition, Vaina *et al.*[84] studied fMRI scans that suggest the recognition of biological motion stimuli may activate brain regions involved in both form and motion recognition. This psychophysical evidence bolsters the idea that automatic face recognition systems should exploit the available temporal information, especially in circumstances involving low resolution videos.

Sequence-based methods draw from these observations insofar as they account for temporal dynamics during recognition. The most popular class, spatiotemporal techniques, combines spatial and temporal cues, ideally improving recognition accuracy in uncontrolled contexts and potentially increasing the efficiency of tracking and recognition. Temporal methods have begun to arise more recently. This group of techniques employs facial movements as identifying biometric characteristics.

6.1. *Spatiotemporal Recognition*

Spatiotemporal recognition schemes model dynamics to estimate identity under the assumptions that idiosyncratic facial motions are accompanied by appearance variations, and, for some algorithms, that identifying movements will be salient. Research on face tracking and face recognition from video has traditionally been performed separately until recent years. Consequently, spatiotemporal recognition methods can be roughly divided into two categories: those that split tracking and recognition into separate tasks performed serially and those that unify tracking and recognition.

Table 7: Selected Split Tracking and Recognition Approach Results. Conditions include indoor/outdoor (i/o), varying pose (p), expression (e), and walking (w). All performance results are given as rank-one recognition rates unless otherwise noted.

Author, Year Face Representation	Subject Count, Video Count, Resolution Recognition Method	Conditions Performance
Liu and Chen,2003[33]	24, 500, 640x480[a]	i,w
PCA vectors	HMMs trained online	98.8%
Aggarwal *et al.*,2004[37]	50, 75, 640x480[b]	i,p
ARMA model	Subspace angle based	90.0%
Mitra *et al.*,2006[86]	55, 165, Not described	e
Frequency domain asymmetry cues	HMMs	96.8%
Hadid and Pietikäinen,2009[32]	24, 96, 640x480[c]	i,w
Extended volume LBP histograms	Chi-square distance	97.9%
-	50, 75, 640x480[b]	i,p
-	-	96.0%

Note:
[a]The authors drew from the CMU MoBo dataset to randomly synthesize 500 sequences.
[b]The authors experimented on the First Honda/UCSD dataset.
[c]Experiments were performed on the original CMU MoBo dataset.

6.1.1. *Split Tracking and Recognition*

Performing tracking and recognition with independent algorithms offers the advantage of flexibility: the tracker is not constrained by the recognition component and vice versa. The most common spatiotemporal approach is to employ a Hidden Markov Model (HMM) for recognition along with a suitable face tracking algorithm. An HMM is comprised by an unobservable Markov chain with a finite number of states connected by transition edges. Markov chains model random processes that proceed through state sequences, such that each state transition is based on the current state and not on the past states. In this context, unobservable means that the state of the model is not directly visible, but outputs and parameters are observable. The states each have an observation probability distribution over the outputs of a modeled system, *e.g.* a face sequence. An HMM can be defined as $\Lambda = (A, B, \pi)$, where A denotes the state transition probability matrix, B expresses the observation probability density functions and π represents the initial state distribution. In face recognition from video applications, a collection of HMMs $\{\Lambda_i\}$ are trained to represent the gallery set. Each Λ_i is trained on the face sequence(s) of a particular individual and thus learns the statistics and dynamics of that person. The identity of a probe sequence observation O is given by the HMM Λ_k for which

$$P(O|\Lambda_k) = \max_i \ P(O|\Lambda_i). \tag{7}$$

In Ref. 33, Liu and Cheng introduce adaptive HMMs to the face recognition from video domain. They apply PCA as a means of dimensionality reduction, *i.e.* the HMMs output sequences of PCA feature vectors. A refinement to the basic HMM learning procedure, wherein the HMM for a recently recognized person learns new sequence information online, provides increased recognition accuracy over time. This adaptive model reached 98.8% recognition accuracy on 500 clips constructed from random subsequences in the CMU MoBo videos. Mitra *et al.*[86] combine a feature representation based on face asymmetry cues from the frequency domain with an HMM set. Their video training and testing data captures the activities of 55 persons displaying three emotions in different clips. An average error rate of 3.3% over 20 trials was achieved. Tistarelli *et al.*[87] went on to employ a two-dimensional generalization of the HMM that incorporates appearance based spatial HMMs, which model the emission distributions of the hidden states in a top-level HMM that characterizes temporal dynamics. To index videos, Eickeler *et al.*[88] apply K-means clustering to group face feature vectors and a collection of two-dimensional HMMs to represent each cluster.

Neural networks provide an alternative means to model state changes while estimating identity over time. For example, Gorodnichy presents an auto-associative neural network framework that accumulates evidence over a series of video frames in Ref. 89. The recognition process begins by passing a pre-processed image into the input layer and determining which output neuron fires subsequent to reaching a stable state. Gorodnichy notes that the neural network layers can incorporate feedback in order to account for temporal dynamics. Barry and Granger employ an evidence accumulator to fuse the results from a fuzzy ARTMAP neural network that performs recognition and an array of Kalman filters that track motion.[90] The individual that most likely resides at a particular position at a specific time is determined by accumulation variables that are updated with the neural network's responses at each time step.

The primary disadvantage of employing HMMs or neural networks is that they implicitly characterize the geometric properties of a moving face, making it difficult to obtain direct estimates of face pose and motion state without special training or separate mechanisms such as the Kalman filter array used in Ref. 90. Aggarwal *et al.*[37] avoid these issues by modeling the face sequence as a linear dynamical system with an autoregressive moving average (ARMA) model. Specifically, the ARMA model is used to characterize changes in appearance due to both pose variation and facial dynamics:

$$x(t+1) = A * x(t) + v(t), \text{and} y(t) = C * x(t) + w(t), \tag{8}$$

where t denotes time, $x(t)$ represents a state vector with components describing

orientation and position, $y(t)$ expresses the observed image, A and C denote matrices that act as linear mappings, and $v(t)$ and $w(t)$ are realizations drawn from separate Gaussian distributions that model noise processes. The model parameters for the noise distributions and the A and C matrices can be efficiently estimated via a closed-form derivation. Once the models are estimated for probe and gallery video sequences alike, ARMA model comparisons are performed with distance metrics based on their subspace angles. 90% recognition accuracy was reached on the First Honda/UCSD dataset.

The HMM (see Refs. 33, 86, 87 and 88), neural network (see Refs. 90 and 89) and ARMA model (see Ref. 37) methods discussed above incorporate global facial features over the sequence. But local information may be equally important to face recognition. It has been found that humans can recognize faces even when a significant number of features cannot be seen. As discussed in Ref. 59 and studied by Sadr *et al.*[91], Davies *et al.*[92] and Fraser *et al.*[93], a single feature such as the eyes or eyebrows is sufficient for people to recognize familiar faces. Similarly, automatic local feature based schemes compensate for pose differences by allowing the geometric configuration between features to be flexible.[94] Robustness to alignment variations can also be achieved via a local approach. Moreover, the spatiotemporal methods discussed above give all dynamic features equal weight and thus idiosyncratic spatiotemporal features do not contribute more to recognition decisions than other facial movements.

Hadid and Pietikäinen increase the influence of identifying spatiotemporal features by incorporating an extended set of volume local binary pattern (LBP) features into a boosting scheme.[32] Hadid and Pietikäinen introduce the volume local binary pattern operator (VLBP), which generalizes the LBP operator that is commonly used to encode spatial information. This is accomplished by treating a face sequence as a rectangular volume and computing the LBP value for each pixel based on a 3-dimensional neighborhood. The VLBP representation only includes a specific number of pixels from three frames at a time and thus fails to incorporate sufficient temporal information or provide flexibility in defining the point neighborhoods. The proposed extended VLBP (EVLBP) operator surmounts these shortcomings by allowing a different numbers of neighboring points to be included from different frames within a temporal window. The AdaBoost learning algorithm determines the best set of EVLBP features, *i.e.* those features that enable the discrimination between subject classes but do not characterize intra-class appearance variations due to expression. Recognition is performed by constructing a vector of local histograms of EVLBP patterns from the selected regions of a probe face sequence and performing nearest neighbor matching with the Chi-square distance metric. These methods achieved respective recognition rates of 97.9% and 96.0% on the CMU Mobo and First Honda/UCSD datasets.

Table 8: Selected Unified Tracking and Recognition Approach Results. Conditions include indoor/outdoor (i/o), varying pose (p), expression (e), and occlusions (c). All performance results are given as rank-one recognition rates unless otherwise noted.

Author, Year Face Representation	Subject Count, Video Count, Resolution Recognition Method	Conditions Performance
Li *et al.*, 2003[95] Identity surfaces	12, 12, Not Described Distance between surfaces	p 100%
Zhou *et al.*, 2004[12] Adaptive appearance model	29, Not Described, 240x360 Probability propagation	i,p,e 100%
Lee *et al.*, 2005,[14] Probabilistic appearance manifolds	15, 30, 640x480[a] Bayesian inference	i,p,c 98.8% max

Note:

[a]The authors used the Second Honda/UCSD dataset.

6.1.2. *Unified Tracking and Recognition*

Kanade and others suggest that proper image registration is critical to pixel-level illumination and pose normalization procedures, as these methods rely on a strong correspondence between the points on the facial surface and the image pixels that cover them.[65] Additionally, matchers that draw from appearance features are particularly sensitive to poor alignment. Using separate tracking and recognition components in a face recognition system makes registration more difficult as the tracker is more likely to return images that do not fit the appearance model used by the recognition algorithm. Results obtained by Lee *et al.*[14] imply that performing tracking and recognition within a common framework results in better alignment and, hence, improved recognition accuracy. The unification of tracking and recognition also enables a seamless integration of pose information into the appearance model, so that score level fusion can be performed in a principled fashion via probability propagation over the sequence.

Li *et al.*[95] introduce the identity surface feature representation that characterizes facial appearance variations due to changes in yaw and tilt. Pose estimates are obtained after fitting a face image to a 3D point distribution and an active appearance model. In turn, these estimates are used transform the face image to a standard view. A nonlinear feature vector is subsequently extracted and used for constructing a piece-wise planar approximation of the identity surface in a kernel discriminant analysis feature space. The distance between identity surfaces is defined as a weighted sum of distances between corresponding tilt and yaw points, while face sequences are treated as trajectories traced out on an identity surface.

Although the trajectory distance accumulates recognition evidence over time, recognition is still deterministic in the identity surface approach. Additional robustness to degraded viewing conditions and outliers can be obtained via a probabilistic approach, such as that proposed in Ref. 38, wherein Zhou *et al.* exploit temporal cues and identity observations to perform face recognition from video with a gallery comprised by still images. Their approach estimates the joint identity and kinematic state distribution over time using evidence acquired over the video frames. To propagate prior joint identity and motion distributions, a particle filter, the CONDENSATION algorithm, is applied. The summed absolute difference between a detected face image and each gallery template is used to calculate the current sample weight for each identity and its associated state. The joint state and identity distribution estimates depend on these weights. Marginalization over the state variable ultimately provides the identity distribution of the current frames. Zhou *et al.*[11] achieve efficiency gains over the computationally intensive CONDENSATION filter by accounting for the discrete nature of the identity variable. Perfect recognition accuracy was achieved on the CMU MoBo dataset when the gallery set contained fast walking and carrying videos or incline and carrying videos and the probe set contained all other videos.

Zhou *et al.*[12] focus on stabilizing the particle filter based tracker with an adaptive appearance model, an adaptive velocity motion model and an adaptive particle count. The adaptive appearance model characterizes the appearances in a face sequence up to a specific time with a mixture of Gaussians. The mixture components represent pair-wise frame changes as well as the stable structure observed over a sequence. Adaptive velocity estimation is performed with a first-order linear prediction method that considers the differences between pairs of consecutive frames. Unlike the methods proposed in Refs. 11 and 38, an adaptive noise variance parameter, which varies with the quality of the motion state prediction, is used to handle large state changes. In addition, the number of particles changes with the noise variance so that fewer particles are used if the noise has a small variance. This feature increases the computational efficiency of the algorithm when the quality of the prediction is high. The use of intra- and extra-personal face spaces along with the adaptive tracker bolsters strong recognition performance: The proposed tracking and recognition framework reached 100% recognition accuracy on a (small) 29-subject database acquired with a hand-held camcorder inside different office environments.

Lee *et al.*[14] propose the probabilistic appearance manifold representation to handle changes caused by pose variation in a robust and transparent fashion. A collection of pose-specific linear subspaces connected by transition probabilities serve as a piecewise approximation of the appearance manifold. Based on this representation, a Bayesian inference framework recursively fuses the recognition results from each frame in a sequence to yield a final decision. Their tracker localizes the subimage of the next frame with the shortest distance to the linear subspace that was nearest to the face in the current frame. In turn, that subimage is passed to

the recognition component.

A manifold learning process clusters faces with similar poses, constructs linear subspaces from the clusters, and approximates the transition probabilities between subspaces. First, the linear subspace learning algorithm performs k-means clustering on the image sequences of a given individual to group images with similar poses. Principal component analysis on each cluster subsequently yields the collection of pose-specific linear subspaces. The temporal continuity inherent to the training sequences is exploited to derive transition probabilities between the linear subspaces. The transition probabilities model the continuity of appearance changes caused by variations in pose; the probability of transitioning between adjacent poses, say from frontal to five degrees to the left, is higher than that of transitioning between distant poses, e.g. 90 degrees right to 90 degrees left.

These methods were evaluated on videos from the Second Honda/UCSD dataset. In videos with and without occlusion, the proposed framework reached 97.8% and 98.8% recognition rates, respectively. The baseline methods performed worse in both contexts. The strongest performing baseline algorithm, a nearest neighbor matcher that operates in the original image space, attained 76.3% and 81.6% recognition rates.

6.1.3. *Temporal Methods*

It has been frequently observed by the face recognition community that changes in expression increase intra-class variance.[32] In other words, expression variations increase the difficultly associated with correctly matching images of a particular person. On the other hand, expressive facial movements can serve as biometric characteristics.

It is well known in psychology that humans use dynamic facial signatures to recognize familiar faces.[96] This notion was corroborated during a study conducted by Lander *et al.*[97], wherein human participants recognized animated face sequences of celebrities more readily than static images. A later study by Lander *et al.*[98] suggested that smiles captured in genuine video sequences aided face recognition more than computer generated smiles, *i.e.* authentic expression changes contain salient identifying characteristics. A battery of tests conducted by Thornton and Kourtzi involving recognition between faces with different expressions indicated that subjects trained with animated sequences, as opposed to static images, had a performance advantage.[99] Further, Piltz *et al.*[100] observed that training subjects with moving faces decreased their reaction times and increased their recognition rates.

It must be acknowledged that the precise underpinnings of these behaviors remain unclear. Nevertheless, these results indicate that automatic face recognition systems can potentially draw identifying information from motion-based characteristics, especially in situations involving the recognition of frequently encountered faces. On a more practical note, temporal cues are not obscured by thick make-up or

Table 9: Selected Temporal Recognition Approach Results. Conditions include indoor/outdoor (i/o) and expression (e). All performance results are given as rank-one recognition rates unless otherwise noted.

Author, Year Face Representation	Subject Count, Video Count, Resolution Recognition Method	Conditions Performance
Benedikt *et al.*,2008[15] Eigen-coefficients	55, 105 3D videos, Not described WDTW	e 99%
Ye and Sim,2010[16] LDPs	11, 66, 640x480 Facial deformation similarity	i,e 30.1% EER

similar facial decorations. These ideas have served as the basis for face recognition research involving temporal features. In contrast to the spatiotemporal methods described above, which incorporate temporal and spatial information during recognition, temporal methods rely solely on dynamic features and thus have attained robustness to expression changes and, in some cases, illumination variations.

For instance, facial motion can be represented by a high-dimensional feature vector obtained from a sequence of dense optical flow fields and compared in terms of the distance between vectors, as proposed by Chen *et al.*[17]. Experimental results on synthetic data featuring subjects speaking two words indicated that this technique has some level of illumination invariance. Matta and Dugelay introduce a temporally oriented algorithm that computes geometrically normalized feature vectors expressing eye, nose and mouth displacements over frame sequences.[101] Benedikt *et al.*[15] employ 3D facial actions as biometric signatures. The feature extraction algorithm performs face model alignment and eigenvector analysis on the 3D face data. The largest eigen-coefficients of the lip regions from the face observations in a sequence represent a subject. In turn, a novel pattern matching technique, Weighted Dynamic Time Warping (WDTW), is used for recognition. WDTW treats a pair of feature vectors as sequences of data points indexed by the video frame number. The WDTW match score accounts for the Euclidean distances between points as well as their first and second derivative differences. The authors found that the utterance of the word "puppy" is a strong behavioral signature as its accompanying facial actions are both distinctive and reproducible. The proposed algorithm reached a 99% rank-one recognition rate on a private dataset comprised by 3D videos of 55 subjects speaking the word "puppy".

The temporal approaches discussed in Refs. 17 and 15 constrain the types of motion that are allowed as subjects must speak particular words, rendering the face recognition process more obtrusive and less practical for uncontrolled applications. Ye and Sim avoid these pitfalls with a strategy that incorporates biometric characteristics drawn from locally similar facial motions, *i.e.* motions that might differ on

a global scale yet share common local features such as the way the muscles stretch or contract around a particular face region.[16] Facial deformation patterns are characterized with the Right-Cauchy Green deformation tensor, which captures how facial features are displaced when the face changes from a neutral expression. The tensors and associated displacement field vectors for each pixel in every frame from a sequence are collected in a set called the Local Deformation Profile (LDP). The associated LDP similarity score measures how similar the deformations are between a pair of faces. Motion similarity is incorporated as a confidence measure to avoid mistaking differences in movement for differences in identity. These techniques were tested on various subsets of the 97 subject Cohn-Kanade database. In a challenging experiment involving recognition across happy, sad, surprise, fear, anger and disgust expressions for 11 subjects, the proposed temporal scheme reached a 30.1% EER. Robustness to thick facial makeup was exhibited in another experiment.

The temporal methods described here are robust to facial decorations and, potentially, illumination. Numerous problems associated with the temporal approach to face recognition from sequences still remain unsolved. One open topic is whether or not facial motions with longer durations allow for more reliable or more accurate recognition. Additionally, the experiments discussed in Refs. 15, 16 and 17 focus on subjects viewed from frontal angles. The extent to which temporal algorithms tolerate out-of-plane rotations remains to be seen. Likewise, the subjects observed in the experimental datasets tend to stand close to the sensor: How quickly does the accuracy of these systems degrade as the resolution decreases? More research is needed before temporal methods can reach the level of maturity of other face recognition approaches.

7. Future Challenges and Directions

Some of the algorithms discussed above have achieved success on challenging datasets. Systems incorporating face recognition technology have already been deployed in surveillance,[102] social networking,[103] and movie indexing domains.[104] Success in these applications is still limited as a number of problems and research challenges remain unsolved or unaddressed. A brief overview of these research issues, potential applications and open problems is given below.

7.1. *Larger and More Challenging Datasets*

The field of face recognition from video lags behind other biometric fields in terms of dataset size. Early work on video-based face recognition used databases containing about 20 subjects.[6] Today, datasets comprised by thousands of videos and hundreds of subjects are available to the public, such as the video collections featured by the Multiple Biometric Grand Challenge.[26] But evaluations on databases of this size are not common. Likewise, large-scale indexing tests are rare in academia, despite the fact that video-indexing systems have a large amount of data readily available from movies, TV shows and web videos. Kim *et al.*[25] have made great strides towards

addressing this problem by aggregating almost 2,000 YouTube videos of almost 50 famous people to form the YouTube Celebrities dataset.

Further, face recognition from video represents a particularly difficult problem due the infinite number of possible appearance variations face sequences can span. The level of difficulty of most current databases is nevertheless lacking. Surveillance quality video datasets that incorporate crowds of people, occlusions, significant amounts of noise and compression artifacts alongside variations in pose, illumination and expression are necessary to evaluate performance in uncontrolled environments. The performance evaluations for sequence-based methods, in particular, have not traditionally drawn from such challenging data. Conversely, the level of difficulty for a dataset should not be so high that it precludes the possibility of researchers making reasonable progress. As methods mature and their robustness to the nuisance factors continues to increase, this situation should improve and the difficulty levels of successive generations of research datasets should continue to increase.

7.2. *The Watch List Task*

In the biometrics domain, the watch list task is more challenging than those of verification and identification. It involves an open set problem and thus requires multiple decisions. The face recognition system must determine whether or not someone is in the gallery and, if that person is found, return a match. Much of the current literature on face recognition from video ignores the watch list application and focuses on the verification and identification instead. However, watch list applications arise often in the law enforcement domain.[3]

7.3. *Clustering Applications*

Research on face clustering has driven the development of video and image indexing software, such as Apple's iPhoto and Google's Picasa, which allows users to automatically organize face image collections. Clustering is the process by which natural groupings or relationships within data are identified. In video-indexing and retrieval applications, clustering can be used to group face images or sequences of the same person together when a database of known identities is not available for matching. Such automatic processing eliminates or reduces the burden of manually labeling thousands or millions of faces in videos or photo albums from personal, movie or news collections.

The initial research on clustering face sequences includes the work of Antonopoulos *et al.*[105], who employ the hierarchical agglomerative clustering algorithm as a means of grouping images of actors from movie videos. Faces are grouped together based on the similarity of scale invariant feature transform descriptors (SIFT). In one of the earliest works on anchor detection, Chan *et al.*[106] propose a method to recognize people who repeatedly appear in news videos that incorporates k-means clustering. Similarly, Raytchev and Murase successfully employ exemplars

to construct high quality face sequence clusters.[107] The learning algorithm proposed therein first tessellates areas in the face space, computes an exemplar face vector for each area, and then alters the clusters to minimize the representation error incurred by the exemplar selections. In Ref. 108, Raytchev and Murase propose a clustering algorithm that iteratively builds a complete graph in which vertices represent frame sequences and edge labels indicate whether connected face sequences most likely correspond to the same person. Clusters are formed from vertex sets on the basis of the edge labels. Raytchev and Murase present another method for clustering that computes forces attraction and repulsion in Ref. 109. With the aim of analyzing internet videos, the automatic content analysis system proposed by Holub *et al.*[110] applies hierarchical agglomerative clustering to organize face image sequences.

Recent work on face clustering from video has begun to directly address nuisance factors, such as pose. Tao and Tan handle the problem of clustering face sequences from movies with significant pose variations by partitioning sequences into subsequences containing faces with similar poses.[111] They present strong results from clustering experiments involving segments from three movies: *Harry Potter and the Goblet of Fire*, *The Queen*, and *Secret*. A similar framework for automatically labeling the faces of characters in TV or movie videos using subtitle and script text is proposed by Sivic *et al.* in Ref. 112. Agglomerative clustering is applied during tracking to link together frontal and profile images of the same face based on the optical flow feature points they share in common; face sequences are compared using person specific multiple kernel discriminative classifiers. Likewise, Yang *et al.*[113,114] associate labels from transcripts with faces and use PCA based face recognition and multi-instance learning to increase video retrieval accuracy despite challenging variations in appearance.

Today, popular commercial applications such as Facebook, Google's Picasa and Apple's iPhoto include face recognition features that attempt to group together images according to identity, ultimately with the goal of reducing the time cost associated with labeling large image collections. As noted by Chellappa *et al.*[65], this application presents unique challenges and opportunities. For instance, algorithms that exploit the overlap between the social networks of multiple users can potentially save people time that would have been spent on labeling the faces of mutual friends. Extending these applications to the video domain presents additional difficulties due to the general lack of constraints and the computational resources involved.

7.4. *Video Understanding Applications*

In domains where the performance on low-level tasks such as identification and tracking is sufficient, more complex face recognition applications can be addressed. High-level tasks, namely analyzing crowds, automatically discovering social groups and determining which individuals appear frequently in a collection videos, constitute but a portion of the future applications of face recognition technology. For example, Vallespi *et al.*[115] propose a meeting understanding system for recognizing

and counting meeting attendees. Face recognition based methods for automatically understanding social network patterns in crowds of people observed by a surveillance camera network are presented by Yu *et al.* in Ref. 116. Barr *et al.*[24] detect individuals that appear unusually often across a set of videos showing related events, with the idea that such individuals might be involved with these activities. The recent emergence of such high-level applications exemplifies the growing trend toward using face recognition from video to understand and track the complementary behaviors of individuals and groups.

7.5. *Mobile Devices*

The ubiquity of mobile phones and tablet devices equipped with digital cameras has introduced a wide range of possible applications for face recognition technology. As of early 2012, Google already provides identity verification software that allows a user to unlock Android devices by presenting his or her face to the camera (see Ref. 117), while Apple recently filed a patent application for an efficient low threshold face recognition pipeline with stages for face detection, verification and basic liveness detection (see Ref. 118).

Likewise, research on face recognition in mobile environments has mostly focused on the verification task. This small body of work has addressed two types of problems, the first of which pertains to imaging conditions and the second of which relates to computational efficiency. Mobile devices typically rely on low quality cameras that often yield noisy and under- or overexposed images. The captured images also tend to be highly compressed to save space. These issues are compounded by the fact that the devices are mobile, which means that the illumination conditions and backgrounds can vary substantially between images of the same face. In regards to efficiency, although many new models of mobile devices incorporate moderate amounts of memory and multi-core processors with clock speeds near or beyond 1 GHz, processing speed is still a key issue since the these devices still lag far behind their immobile counterparts on the desktop or in the server room. As far back as 2005, Venkataramani *et al.* [119] studied methods for coping with the low quality imagery acquired by mobile cameras. Hadid *et al.*[120] presented a simple yet effective detection and verification scheme that uses the Viola-Jones method to find faces and a local binary pattern matcher for authentication in 2007. This system was able to verify faces at a rate of two frames per second on a smart phone equipped with a 220 MHz ARM 9 processor.

Numerous open issues and new applications remain. From a practical standpoint, verification on mobile devices is a relatively easy task, as users generally will cooperate by presenting their face at an upright, frontal pose within a short distance from the camera. False negatives and acquisition failures do not incur a significant cost as users can fall back to the traditional PIN based authentication interface. A much more difficult problem is to analyze faces in arbitrary photographs and automatically identify people by treating personal photo albums as *ad hoc* galleries.

More efficient methods for detection, registration and identification will be required before face recognition can be applied in such contexts. For instance, little work has been done on adapting robust face trackers to find faces in mobile environments at arbitrary distances from the camera. Face trackers are generally more efficient than detectors since they do not perform a full search over all of the frames, and they can potentially improve registration.[14] The potential research directions also extend into the domains of social network analysis and law enforcement.

7.6. *Multimodal Approaches*

Much of the research on face recognition from video has focused on representing individuals in terms of the appearance, structure or dynamics of their faces. On the other hand, a variety of identifying characteristics are typically visible and complementary forms of information such as audio often accompany videos. Information from the face can potentially be fused with that from other biometric modalities to increase recognition accuracy or to compensate for scenarios where some of the sources cannot be observed. The improvements in performance are proportional to how strongly the various modalities are correlated. The human face notwithstanding, possible biometric modalities include gait, voice, typing style, signature and the iris, amongst others. As an indicator of the current state of multimodal recognition from video, a small selection of recent works that incorporate face sequences with other data sources are briefly discussed below.

Information fusion approaches can either be hierarchical, holistic, or a combination thereof. Hierarchical methods employ different algorithms for distinct modalities at different times. The algorithms that execute later use information from algorithms that complete earlier. For instance, Chellappa *et al.*[121] propose the use of view-invariant gait recognition in scenarios where an individual is located far from the camera. The gait recognition results are used to narrow the search space for a face recognition algorithm that operates when the individual nears the camera. The face recognition process thus becomes more efficient as fewer face comparisons are required. Identification can also occur over a wider range of conditions because the operating ranges of each recognition component offset one another.

Holistic methods fuse the match scores, decisions or data from multiple information source. In Ref. 121, Chellappa *et al.* employ score level fusion over the gait and face modalities. Pentland *et al.*[122] combine speaker identification and face recognition with a Bayesian network, while Weng *et al.*[123] introduce the incremental hierarchical discriminating regression (IHDR) tree and use it to map faces and audio clips to identity labels. Song *et al.*[124] employ a variant of a multiple-instance learning algorithm along with automatic speech recognition to construct face appearance models.

7.7. *Temporal Feature Aging*

Security and surveillance systems that compare video sequences acquired over long periods will naturally benefit from biometric characteristics that are invariant to age. The alternative of updating the gallery on a regular basis can require a significant number of man hours over time. This fact, amongst many others, has motivated research on face aging (see Ref. 125 for a review). Research on aging and its effects on face recognition from video has been neglected despite the growing interest in face aging. A multitude of questions remain unaddressed in this area:

- How and to what extent do the ways in which people make certain expressions change as they age?
- Do some facial regions move in ways that are easier to recognize over time relative to other regions?
- Are spatial features more robust to aging than temporal features?
- Can automatic spatiotemporal feature aging be performed to mitigate the effects stemming from age differences?
- Conversely, do temporal features capture information which can be used to predict the age of a person?
- What types of movement should be captured in video datasets to test hypotheses about aging?

The interplay between facial dynamics and aging effects presents a rich variety of open problems.

7.8. *Sparse Representation*

Recent developments in the theory of compressive sensing and sparse representation have played an increasingly large role in many research disciplines, including face recognition.[126,127,128] Compressive sensing is a reconstruction technique for generating a signal such as a face pattern from an overcomplete basis. The underlying assumption is that the most useful reconstruction is sparse in that it should only depend on a small number of basis vectors and the corresponding coefficient vector should largely consist of values near zero. In the case of face recognition, a newly observed face pattern is reconstructed using a linear combination of the training patterns. This problem is generally underdetermined because the dimensionality of the input data typically exceeds the size of the training set, *i.e.* the pattern can be reconstructed with multiple coefficient vectors. However, a sufficiently sparse coefficient vector can compactly represent the test pattern and generally has non-zero entries for the training patterns of one class. Such a coefficient vector thus indicates the class of the test pattern. The classification problem is consequently reduced to the problem of computing a sufficiently sparse coefficient vector, which can be solved in polynomial time with respect to the number of training samples using a variety of $L1$-minimization algorithms.

The body of work on sparse representation techniques for image based face

recognition is growing. In Ref. 126, Wright *et al.* propose the face recognition framework discussed above and show that it is robust to noise and occlusion, especially when the face patterns consist of pixel features. Later work by Wagner *et al.*[127] extends this scheme by increasing its robustness to illumination variations and alignment errors. Liao *et al.*[128] avoid the alignment problem altogether while improving robustness to pose changes by representing faces with SIFT descriptors. They formulate the reconstruction problem in terms of the descriptors found on test images and a basis of training image descriptors. Sparse representation methods have also been evaluated on tracking problems. A vehicle tracker that employs a particle filter for motion state estimation along with sparse representation techniques for recognition is introduced by Mei *et al.* in Ref. 129.

The inherent robustness to occlusion and noise potentially makes sparse representation techniques ideal for identifying faces in unconstrained situations or with non-cooperative subjects. Conversely, solving an $L1$-minimization problem is more computationally intensive than invoking a nearest neighbor classifier. This difference currently renders real-time recognition infeasible. Refs. 126 and 127 report average recognition times that are on the order of seconds for desktop PCs running experiments on thousands of training images. The appearance based recognition algorithms, Refs. 126, 127 and 129, downsample images to compensate for the computational burden of their approaches. Wagner *et al.*[127] suggest that their alignment algorithm could easily be extended to form an efficient multiscale scheme. Along the same lines, Liao *et al.*[128] tradeoff accuracy for speed by filtering the dictionary of SIFT descriptors before reconstruction. An average recognition time of 0.8 seconds was reached on server grade hardware with no loss in accuracy when the dictionaries were restricted to 1,000 descriptors. Additional efficiency gains are required before sparse representation techniques can be employed in real-time. Moreover, the applicability of the sparseness assumption to temporal and spatiotemporal face representations is still unknown.

7.9. *The Spatiotemporal Tradeoff*

The sequence-based approach exploits the temporal continuity inherent to videos and so handles degraded viewing conditions well. In Ref. 48, Hadid and Pietikäinen present a small scale comparison of simple set-based algorithms to popular spatiotemporal algorithms on low resolution face images. Two set-based methods that combine sum fusion with PCA and LDA matchers were evaluated alongside the ARMA and HMM spatiotemporal algorithms on the CMU MoBo dataset. The authors downsampled the face images to a variety of resolutions, the lowest of which was 10x10. The PCA, LDA, ARMA and HMM based algorithms reached respective recognition rates of 60.6%, 56.5%, 71.2% and 74.2% on the 10x10 images. These results suggest that these particular set-based techniques require relatively high resolution images in order to achieve comparable accuracy. Conversely, it was found that the spatiotemporal algorithms often need to observe a significant number of

frames before they can offer stable recognition results. The emperical results presented in Ref. 48 indicate that HMMs may need to observe 200 or more video frames, or about six seconds worth of footage for a 30 frame per second video, before they can surpass the recognition performance of the simple PCA and LDA algorithms. Likewise, the performance of the ARMA model also suffered before a sufficient number of frames passed. This study begins to answer the question: What are the inherent trade-offs associated with relying on temporal information more than spatial cues and vice versa?

Psychological observations suggest that humans strike a balance between processing spatial and temporal evidence when recognize faces. Burton *et al.*[130] observed that human subjects recognize faces more accurately in poor quality surveillance video when the recorded individuals are familiar colleagues instead of people that they do not encounter often. As to how dynamics aid recognition, cues from expressive or speech related movements have been shown to aid recognition more than rigid motion alone.[131] This phenomenon suggests that humans obtain identifying behavioral characteristics from face sequences as opposed to cues about the 3D structure of the face. Even though people rely on static appearance cues more so than dynamic characteristics, especially when recognizing unfamiliar faces,[5,83] results obtained by Davies *et al.*[132] show that images that only contain high frequency content in the form of unshaded edges are exceptionally difficult to recognize.

Addressing this question in the field of automatic face recognition from video will require a larger scale study than that presented in Ref. 48. Comparing the sophisticated set-based techniques such as the super-resolution schemes presented in Refs. 54, 53, 49, 57 and 58 might give a better indication of the relative performance merits of drawing from temporal versus spatial information. Likewise, comparisons against 3D model based methods, *e.g.* those discussed in Refs. 27, 60, 29 and 28, could provide insight into the trade-offs associated with using geometric characteristics. Such results would benefit the fields of automatic and human based face recognition alike.

8. Acknowledgments

This research is supported by the Central Intelligence Agency, the Biometrics Task Force and the Technical Support Working Group through US Army contract W91CRB-08-C-0093. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of our sponsors.

References

1. BKA, "Final Report Foto-Fahndung." [Online]. Available: http://www.bka.de
2. J. Phillips, "Video Challenge Problem Multiple Biometric Grand Challenge: Preliminary Results of Version 1," 2008.
3. P. J. Phillips, P. Grother, R. Micheals, D. M. Blackburn, E. Tabassi, and M. Bone,

"Face Recognition Vendor Test 2002," *Proc. IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, p. 44, 2003.
4. S. K. Zhou and R. Chellappa, *Beyond a Single Still Image: Face Recognition from Multiple Still Images and Videos.* Academic Press, 2005.
5. A. J. O'Toole, D. A. Roark, and H. Abdi, "Recognizing Moving Faces: A Psychological and Neural Synthesis," *Trends in Cognitive Sciences*, pp. 261–266, 2002.
6. W. Zhao, R. Chellappa, P. J. Phillips, and A. Rosenfeld, "Face Recognition: A Literature Survey," *ACM Computing Surveys*, vol. 35, no. 4, pp. 399–458, 2003.
7. P. Phillips, J. Beveridge, B. Draper, G. Givens, A. O'Toole, D. Bolme, J. Dunlop, Y. M. Lui, H. Sahibzada, and S. Weimer, "An Introduction to the Good, the Bad, the Ugly Face Recognition Challenge Problem," *Proc. 2011 IEEE Conference on Automatic Face and Gesture Recognition*, pp. 346–353, 2011.
8. C. Sanderson and K. K. Paliwal, "On the Use of Speech and Face Information for Identity Verification," *IDIAP*, 2004.
9. A. Ross, "An Introduction to Multibiometrics," *Proc. 15th European Signal Processing Conference (EUSIPCO)*, 2007.
10. L. Baoxin and R. Chellappa, "A Generic Approach to Simultaneous Tracking and Verification in Video," *IEEE Transactions on Image Processing*, vol. 11, no. 5, pp. 530–544, May 2002.
11. S. Zhou, V. Krueger, and R. Chellappa, "Probabilistic Recognition of Human Faces from Video," *Computer Vision and Image Understanding*, vol. 91, pp. 214–245, 2003.
12. S. Zhou, R. Chellappa, and B. Moghaddam, "Visual Tracking and Recognition Using Appearance-Adaptive Models in Particle Filters," *IEEE Transactions on Image Processing*, vol. 13, pp. 1434–1456, 2004.
13. K. Lee, J. Ho, M. Yang, and D. Kriegman, "Video-based Face Recognition Using Probabilistic Appearance Manifolds," vol. 1, pp. I–313 – I–320, 2003.
14. ——, "Visual Tracking and Recognition Using Probabilistic Appearance Manifolds," *Computer Vision and Image Understanding*, vol. 99, no. 3, pp. 303–331, 2005.
15. L. Benedikt, D. Cosker, P. Rosin, and D. Marshall, "3D Facial Gestures in Biometrics: from Feasibility Study to Application," *Proc. 2008 IEEE International Conference on Biometrics: Theory, Applications and Systems*, pp. 1–6, 2008.
16. N. Ye and T. Sim, "Towards General Motion-based Face Recognition," *Proc. 2010 IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2598–2605, 2010.
17. L. Chen, H. Liao, and J. Lin, "Person Identification Using Facial Motion," *Proc. 2001 International Conference on Image Processing*, vol. 2, pp. 677–680, 2001.
18. H. Wang, Y. Wang, and Y. Cao, "Video-based Face Recognition: A Survey," *World Academy of Science, Engineering and Technology*, vol. 60, pp. 293–302, 2009.
19. R. Goh, L. Liu, X. Liu, and T. Chen, "The CMU Face In Action (FIA) Database," *Proc. Analysis and Modeling of Faces and Gestures*, pp. 255–263, 2005.
20. R. Gross and J. Shi, "The CMU Motion of Body (MoBo) Database," Robotics Institute, Tech. Rep. CMU-RI-TR-01-18, June 2001.
21. "University of Cambridge Face Database." [Online]. Available: http://mi.eng.cam.ac.uk/~oa214/academic/
22. "Faces96 database." [Online]. Available: http://cswww.essex.ac.uk/mv/allfaces/faces96.html
23. C. Sanderson, "Biometric Person Recognition: Face, Speech and Fusion," 2008. [Online]. Available: http://www.itee.uq.edu.au/~conrad/vidtimit/
24. J. R. Barr, K. W. Bowyer, and P. J. Flynn, "Detecting Questionable Observers Using Face Track Clustering," *Proc. 2011 IEEE Workshop on Applications of Computer*

Vision, pp. 182–189, 2011.
25. M. Kim, S. Kumar, V. Pavlovic, and H. Rowley, "Face Tracking and Recognition with Visual Constraints in Real-World Videos," *Proc. 2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1 –8, 2008.
26. J. Phillips, "Video Challenge Problem Multiple Biometric Grand Challenge Preliminary: Results of Version 2," 2009.
27. U. Park and A. Jain, "3D Model-Based Face Recognition in Video," *Proc. 2nd International Conference on Biometrics*, pp. 1085–1094, 2007.
28. X. Liu and T. Chen, "Face Mosaicing for Pose Robust Video-Based Recognition," *Proc. of the 8th Asian Conference on Computer Vision*, pp. 662–671, 2007.
29. U. Park, A. Jain, and A. Ross, "Face Recognition in Video: Adaptive Fusion of Multiple Matchers," *Proc. 2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, June 2007.
30. R. Wang, S. Shan, X. Chen, and W. Gao, "Manifold-Manifold Distance with Application to Face Recognition Based on Image Set," *Proc. 2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.
31. A. Hadid and M. Pietikäinen, "Selecting Models from Videos for Appearance-Based Face Recognition," *Proc. 17th International Conference on Pattern Recognition*, vol. 1, pp. 304–308, 2004.
32. ——, "Combining Appearance and Motion for Face and Gender Recognition from Videos," *Pattern Recognition*, vol. 42, no. 11, pp. 2818–2827, 2009. [Online]. Available: http://www.sciencedirect.com/science/article/B6V14-4VT5TH4-1/2/bbffafdaea6aefd375c012104c0bbfb7
33. X. Liu and T. Cheng, "Video-based Face Recognition Using Adaptive Hidden Markov Models," *Proc. 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 340–345, 2003.
34. S. K. Zhou and R. Chellappa, "From Sample Similarity to Ensemble Similarity: Probabilistic Distance Measures in Reproducing Kernel Hilbert Space," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 917–929, 2006.
35. A. Mian, "Unsupervised Learning from Local Features for Video-Based Face Recognition," *Proc. 2008 IEEE International Conference on Automatic Face and Gesture Recognition*, p. 6, 2008.
36. D. Thomas, K. W. Bowyer, and P. J. Flynn, "Multi-frame Approaches To Improve Face Recognition," *Proc. 2007 IEEE Workshop on Motion and Video Computing*, p. 19, 2007.
37. G. Aggarwal, A. Chowdhury, and R. Chellappa, "A System Identification Approach for Video-Based Face Recognition," *Proc. 2004 International Conference on Pattern Recognition*, vol. 4, pp. 175–178, 2004.
38. S. Zhou, V. Krueger, and R. Chellappa, "Face Recognition from Video: A CONDENSATION Approach," *Proc. 2002 IEEE Conference on Automatic Face and Gesture Recognition*, pp. 221–226, 2002.
39. P. Viola and M. Jones, "Robust Real-Time Face Detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.
40. "IEEE Computer Vision and Pattern Recognition 2011." [Online]. Available: http://www.cvpr2011.org/
41. H. A. Rowley, S. Baluja, and T. Kanade, "Neural Network-Based Face Detection," *IEEE Transactions On Pattern Analysis and Machine intelligence*, vol. 20, pp. 23–38, 1998.
42. E. Hjelmás and B. K. Low, "Face Detection: A Survey," *Computer Vision and Image Understanding*, vol. 83, no. 3, pp. 236–274, 2001.

43. M. H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting Faces in Images: A Survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 1, pp. 34–58, 2002.
44. D. Comaniciu, V. Ramesh, and P. Meer, "Real-Time Tracking of Non-Rigid Objects Using Mean Shift," *Proc. 2000 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, p. 2142, 2000.
45. Y. Xu and A. Roy-Chowdhury, "Integrating Motion, Illumination, and Structure in Video Sequences with Applications in Illumination-Invariant Tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, pp. 793–806, 2007.
46. A. Azarbayejani and A. Pentland, "Recursive Estimation of Motion, Structure, and Focal Length," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 6, pp. 562–575, 1995.
47. P. Kakumanu, S. Makrogiannis, and N. Bourbakis, "A Survey of Skin-Color Modeling and Detection Methods," *Pattern Recognition*, vol. 40, no. 3, pp. 1106–1122, 2007. [Online]. Available: http://www.sciencedirect.com/science/article/B6V14-4KJDWRF-2/2/0897784a8697ffcd6653a1aa639c8ec6
48. A. Hadid and M. Pietikäinen, "An Experimental Investigation about the Integration of Facial Dynamics in Video-Based Face Recognition," *Electronic Letters on Computer Vision and Image Analysis*, vol. 5, no. 1, pp. 1–13, 2005.
49. B. Gunturk, A. Batur, Y. Altunbasak, M. Hayes, and R. Mersereau, "Eigenface-domain Super-resolution for Face Recognition," *IEEE Transactions on Image Processing*, pp. 597–606, 2003.
50. S. Farsiu, D. Robinson, M. Elad, and P. Milanfar, "Advances and Challenges in Super-Resolution," *International Journal of Imaging Systems and Technology*, vol. 14, no. 2, pp. 47–57, 2004. [Online]. Available: http://dx.doi.org/10.1002/ima.20007
51. S. Park, M. Park, and M. Kang, "Super-resolution Image Reconstruction: A Technical Overview," *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 21–36, 2003.
52. W. Freeman, T. Jones, and E. Pasztor, "Example-based Super-resolution," *IEEE Computer Graphics and Applications*, vol. 22, no. 2, pp. 56–65, 2002.
53. O. Arandjelović and R. Cipolla, "A Manifold Approach to Face Recognition from Low Quality Video Across Illumination and Pose using Implicit Super-Resolution," *Proc. 2007 IEEE International Conference on Computer Vision*, 2007.
54. M. Al-Azzeh, A. Eleyan, and H. Demirel, "PCA-based Face Recognition from Video Using Super-resolution," *Proc. 2008 International Symposium on Computer and Information Sciences*, pp. 1–4, 2008.
55. W. Freeman, T. Jones, and E. Pasztor, "Improving Resolution by Image Registration," *CVGIP: Graphical Models and Image Processing*, vol. 53, pp. 231–239, 1991.
56. X. Zhou and B. Bhanu, "Human Recognition Based on Face Profiles in Video," *Proc. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, p. 15, June 2005.
57. R. R. Jillela and A. Ross, "Adaptive Frame Selection for Improved Face Recognition in Low-resolution Videos," *Proc. 2009 International Joint Conference on Neural Networks*, pp. 2835–2841, 2009.
58. S. Baker and T. Kanade, "Limits on Super-resolution and How to Break Them," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1167–1183, 2002.
59. P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell, "Face Recognition by Humans: Nineteen Results All Computer Vision Researchers Should Know About," *Proceed-*

ings of the IEEE, vol. 94, no. 11, pp. 1948–1962, 2006.
60. Y. Xu, A. Roy-Chowdhury, and K. Patel, "Integrating Illumination, Motion, and Shape Models for Robust Face Recognition in Video," *EURASIP Journal on Advances in Signal Processing*, 2008.
61. U. Park, H. Chen, and A. Jain, "3D Model-Assisted Face Recognition in Video," *Proc. 2005 Canadian Conference on Computer and Robot Vision*, pp. 322–329, 2005.
62. D. Thomas, K. Bowyer, and P. Flynn, "Multi-Factor Approach to Improving Recognition Performance in Surveillance-Quality Video," pp. 1–7, October 2008.
63. D. P. Robertson and R. Cipolla, *Practical Image Processing and Computer Vision*. John Wiley, 2009, ch. Structure from Motion.
64. V. Krüger, R. Gross, and S. Baker, "Appearance-Based 3-D Face Recognition from Video," *Proc. 2002 German Symposium on Pattern Recognition Symposium*, pp. 566–574, 2002.
65. Chellappa, R. and Sinha, P. and Phillips, P. J., "Face Recognition by Computers and Humans," *Computer*, vol. 43, pp. 46–55, 2010.
66. A. Hadid and M. Pietikäinen, "Manifold Learning for Video-to-Video Face Recognition," *Biometric ID Management and Multimodal Communication*, pp. 9–16, 2009.
67. O. Yamaguchi, K. Fukui, and K. Maeda, "Face Recognition Using Temporal Image Sequence," *Proc. 1998 IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 318–323, 1998.
68. G. Shakhnarovich, J. W. Fisher, III, and T. Darrell, "Face Recognition from Long-Term Observations," *Proc. 2002 European Conference on Computer Vision-Part III*, pp. 851–868, 2002.
69. W. Fan and D. Yeung, "Locally Linear Models on Face Appearance Manifolds with Application to Dual-Subspace Based Classification," *Proc. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 1384–1390, 2006.
70. O. Arandjelović and R. Cipolla, "An Information-Theoretic Approach to Face Recognition from Face Motion Manifolds," *Image and Vision Computing*, vol. 24, no. 6, pp. 639–647, 2006.
71. ——, "A Methodology for Rapid Illumination-Invariant Face Recognition Using Image Processing Filters," *Computer Vision and Image Understanding*, vol. 113, no. 2, pp. 159–171, 2009.
72. Lina, T. Takahashi, I. Ide, and H. Murase, "Incremental Unsupervised-Learning of Appearance Manifold with View-Dependent Covariance Matrix for Face Recognition from Video Sequences," *IEICE Transactions on Information and Systems*, vol. E92.D, no. 4, pp. 642–652, 2009.
73. S. T. Roweis and L. K. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science*, vol. 290, pp. 2323–2326, 2000.
74. W. Fan and D. Yeung, "Face Recognition with Image Sets Using Hierarchically Extracted Exemplars from Appearance Manifolds," *Proc. 2006 IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 177–82, 2006.
75. P. Turaga, A. Veeraraghavan, and R. Chellappa, "Statistical Analysis on Stiefel and Grassmann Manifolds with Applications in Computer Vision," *Proc. 2008 Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.
76. Y. Zhang and A. M. Martínez, "A Weighted Probabilistic Approach to Face Recognition from Multiple Images and Video Sequences," *Image and Vision Computing*, vol. 24, no. 6, pp. 626–638, 2006.
77. J. Stallkamp, H. K. Ekenel, and R. Stiefelhagen, "Video-based Face Recognition on Real-World Data," *Proc. 2007 IEEE International Conference on Computer Vision*,

vol. 206, pp. 1–8.
78. M. Hubert, P. J. Rousseeuw, and K. V. Branden, "ROBPCA: a New Approach to Robust Principal Component Analysis," *Technometrics*, vol. 47, pp. 64–79, 2005.
79. S. Berrani and C. Garcia, "Enhancing Face Recognition from Video Sequences Using Robust Statistics," *Proc. 2005 IEEE Conference on Advanced Video and Signal Based Surveillance*, pp. 324–329, September 2005.
80. J. B. Tenenbaum and J. C. Silva, V. d. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," *Science*, vol. 290, pp. 2319–2323, 2000.
81. "Cognitec — The Face Recognition Company: FaceVACS at a glance." [Online]. Available: http://www.cognitec-systems.de/Technology.11.0.html
82. Q. Xiong and C. Jaynes, "Mugshot Database Acquisition in Video Surveillance Networks Using Incremental Auto-Clustering Quality Measures," *Proc. 2003 IEEE Conference on Advanced Video and Signal Based Surveillance*, p. 191, 2003.
83. B. Knight and A. Johnston, "The Role of Movement in Face Recognition," *Visual Cognition*, vol. 4, no. 3, pp. 265–273, September 1997.
84. L. M. Vaina, J. Solomon, S. Chowdhury, P. Sinha, and J. W. Belliveau, "Functional Neuroanatomy of Biological Motion Perception in Humans," *Proc. 2001 National Academy of Sciences of the United States of America*, vol. 98, no. 20, pp. 11 656–11 661, 2001. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=58785
85. B. Knappmeyer, I. M. Thornton, and H. Bülthoff, "Facial Motion Can Determine Facial Identity," *Journal of Vision*, vol. 1, no. 3, 2001.
86. S. Mitra, M. Savvides, and B. V. K. V. Kumar, "Human Face Identification from Video Based on Frequency Domain Asymmetry Representation Using Hidden Markov Models ," *Lecture Notes in Computer Science: Multimedia Content Representation, Classification and Security*, pp. 26–33, 2006.
87. M. Tistarelli, M. Bicego, and E. Grosso, "Dynamic Face Recognition: From Human to Machine Vision," *Image and Vision Computing*, vol. 27, pp. 222–32, 2009.
88. S. Eickeler, F. Wallhoff, U. Lurgel, and G. Rigoll, "Content Based Indexing of Images and Video Using Face Detection and Recognition Methods," *Proc. 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 1505–1508 vol.3, 2001.
89. D. Gorodnichy, "Video-based Framework for Face Recognition in Video," *Proc. 2005 Canadian Conference on Computer and Robot Vision*, pp. 330 – 338, 9-11 2005.
90. M. Barry and E. Granger, "Face Recognition in Video Using a What-and-Where Fusion Neural Network," *Proc. 2007 International Joint Conference on Neural Networks*, pp. 2256–2261, 2007.
91. Sadr, J., Jarudi, I. and Sinha, P., "The Role of Eyebrows in Face Recognition," *Perception*, vol. 32, pp. 285–293, 2003.
92. G. Davies, H. Ellis, and J. Shepherd, "Cue Saliency in Faces as Assessed by the Photofit Technique," *Perception*, vol. 6, pp. 263–269, 1977.
93. I. Fraser, G. Craig, and D. Parker, "Reaction Time Measures of Feature Saliency in Schematic Faces," *Perception*, vol. 19, no. 5, pp. 661–673, 1990.
94. B. Heisele, P. Ho, J. Wu, and T. Poggio, "Face Recognition: Component Based versus Global Approaches," *Computer Vision and Image Understanding*, vol. 91, pp. 6–21, 2003.
95. Y. Li, S. Gong, and H. Liddell, "Constructing Facial Identity Surfaces for Recognition," *International Journal of Computer Vision*, vol. 53, no. 1, pp. 71–92, 2003.
96. D. A. Roark, S. E. Barrett, M. Spence, H. Abdi, and A. J. O'Toole, "Memory for Moving Faces: Psychological and Neural Perspectives on the Role of Motion in Face

Recognition," *Behavioral and Cognitive Neuroscience Reviews*, vol. 2, pp. 15–46, 2003.
97. K. Lander, F. Christie, and V. Bruce, "The Role of Movement in the Recognition of Famous Faces," *Memory & Cognition*, vol. 27, pp. 974–985, 1999.
98. K. Lander, L. Chuang, and L. Wickham, "Recognizing Face Identity from Natural and Morphed Smiles," *The Quarterly Journal of Experimental Psychology*, vol. 59, pp. 801–808, 2006.
99. I. M. Thornton and Z. Kourtzi, "A Matching Advantage for Dynamic Human Faces," *Perception*, vol. 31, pp. 113–132, 2002.
100. K. S. Pilz, I. M. Thornton, and H. H. Bülthoff, "A Search Advantage for Faces Learned in Motion," *Experimental Brain Research*, vol. 171, pp. 436–447, 2006.
101. F. Matta and J.-L. Dugelay, "Video Face Recognition: A Physiological and Behavioural Multimodal Approach," *Proc. 2007 IEEE International Conference on Image Processing*, vol. 4, pp. IV–497 – IV–500, 2007.
102. Cognitec Systems GmbH, "Video Surveillance Systems with Face Recognition Technology." [Online]. Available: http://www.security-technologynews.com/article/video-surveillance-systems-with-face-recognition-technology.html
103. H. Popkin, "Facebook's Facial Recognition Knows Who Your Friends Are?" *MSNBC.com*, Dec. 2010.
104. K. Dawson, "Analyzing (All of) Star Trek with Face Recognition," *Slashdot*, 2009.
105. P. Antonopoulos, N. Nikolaidis, and I. Pitas, "Hierarchical Face Clustering Using SIFT Image Features," *Proc. 2007 IEEE Symposium on Computational Intelligence in Image and Signal Processing*, pp. 325–329, 2007.
106. Y. Chan, S. Lin, Y. Tan, and S. Kung, "Video Shot Classification Using Human Faces," *Proc. 1996 IEEE International Conference on Image Processing*, vol. 3, pp. 843–6, 1996.
107. B. Raytchev and H. Murase, "VQ-faces - Unsupervised Face Recognition from Image Sequences," *Proc. 2002 IEEE International Conference on Image Processing*, vol. 2, pp. II–809–12, 2002.
108. ——, "Unsupervised Face Recognition by Associative Chaining," *Pattern Recognition*, vol. 36, pp. 245–57, 2003.
109. ——, "Unsupervised Face Recognition from Image Sequences Based on Clustering with Attraction and Repulsion," *Proc. 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. II–25–30, 2001.
110. A. Holub, P. Moreels, A. Islam, A. Makhanov, and R. Yang, "Towards Unlocking Web Video: Automatic People Tracking and Clustering," *Proc. 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 47–54, 2009.
111. J. Tao and Y.-P. Tan, "Efficient Clustering of Face Sequences with Application to Character-Based Movie Browsing," *Proc. 2008 IEEE International Conference on Image Processing* , pp. 1708–1711, Oct. 2008.
112. J. Sivic, M. Everingham, and A. Zisserman, " 'Who are you?' - Learning Person Specific Classifiers from Video," *Proc. 2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
113. J. Yang, M.-y. Chen, and A. Hauptmann, "Finding Person X: Correlating Names with Visual Appearances ," *Proc. 2004 International Conference on Image and Video Retrieval*.
114. J. Yang, R. Yan, and A. G. Hauptmann, "Multiple Instance Learning for Labeling Faces in Broadcasting News Video," *Proc. 2005 ACM International Conference on Multimedia*, pp. 31–40, 2005.
115. C. Vallespi, F. De la Torre, M. Veloso, and T. Kanade, "Automatic Clustering of

Faces in Meetings," *Proc. 2006 IEEE International Conference on Image Processing*, pp. 1841–1844, 2006.
116. T. Yu, S.-N. Lim, K. Patwardhan, and N. Krahnstoever, "Monitoring, Recognizing and Discovering Social Networks," *Proc. 2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1462–1469, 2009.
117. Google Inc., "Introducing Android 4.0." [Online]. Available: http://www.android.com/about/ice-cream-sandwich/
118. Apple Inc., "Low Threshold Face Recognition," *United States Patent Application 20110317872*.
119. K. Venkataramani, S. Qidwai, and B. Vijayakumar, "Face Authentication from Cell Phone Camera Images with Illumination and Temporal Variations," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 35, no. 3, pp. 411–418, 2005.
120. A. Hadid, J. Heikkila, O. Silven, and M. Pietikainen, "Face and Eye Detection for Person Authentication in Mobile Phones," *Proc. 2007 ACM/IEEE Conference on Distributed Smart Cameras*, pp. 101–108, 2007.
121. R. Chellappa, A. K. Roy Chowdhury, and A. Kale, "Human Identification Using Face and Gait," *Multimodal Surveillance: Sensors, Algorithms and Systems*, 2007.
122. A. Pentland, T. Jebara, B. Clarkson, and T. Choudury, "Multimodal Person Recognition Using Unconstrained Audio and Video," *Proc. 1999 Audio-and Video-based Biometric Person Authentication*, pp. 176–181, 1999.
123. J. Weng, C. H. Evans, and W.-S. Hwang, "An Incremental Learning Method for Face Recognition under Continuous Video Stream," *Proc. 2000 IEEE International Conference on Automatic Face and Gesture Recognition*, p. 251, 2000.
124. S. Xiaodan, C.-Y. Lin, and M.-T. Sun, "Cross-Modality Automatic Face Model Training from Large Video Databases," *Proc. 2004 Computer Vision and Pattern Recognition Workshop*, p. 91, 2004.
125. N. Ramanathan, R. Chellappa, and S. Biswas, "Computational Methods for Modeling Facial Aging: A Survey," *Journal of Visual Languages & Computing*, vol. 20, no. 3, pp. 131–144, 2009.
126. J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust Face Recognition via Sparse Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.
127. A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi, and Y. Ma, "Toward a Practical Face Recognition System: Robust Alignment and Illumination by Sparse Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 2, pp. 372–386, 2012.
128. S. Liao and A. Jain, "Partial Face Recognition: An Alignment Free Approach," *Proc. 2011 IEEE International Joint Conference on Biometrics*, pp. 1–8, 2011.
129. X. Mei and H. Ling, "Robust Visual Tracking and Vehicle Classification via Sparse Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence* , vol. 33, no. 11, pp. 2259–2272, 2011.
130. M. Burton, S. Wilson, M. Cowan, and V. Brucen, "Face Recognition in Poor-Quality Video: Evidence from Security Surveillance," *Psychological Science*, vol. 10, no. 3, pp. 243–248, 1999.
131. K. Lander and L. Chuang, "Why are Moving Faces Easier to Recognize?" *Visual Cognition*, 2005.
132. G. Davies, H. Ellis, and J. Shepherd, "Face Recognition Accuracy as a Function of Mode of Representation," *Journal of Applied Psychology*, vol. 63, pp. 180–187, 1978.

Jeremiah R. Barr received his B.S. in Computer Science from Mount Mercy University and is currently pursuing a Ph.D. in Computer Science at the University of Notre Dame. His research interests include biometrics, computer vision, pattern recognition and machine learning.

Kevin W. Bowyer currently serves as Schubmehl-Prein Professor and Chair of the Department of Computer Science and Engineering at the University of Notre Dame. His recent research activities focus on problems in biometrics and in data mining. Professor Bowyer has served as Editor-In-Chief of the IEEE Transactions on Pattern Analysis and Machine Intelligence, recognized as the premier journal in its areas of coverage, and as EIC of the IEEE Biometrics Compendium, the IEEEs first virtual journal. Professor Bowyer was elected an IEEE Fellow for his research in object recognition. Professor Bowyer was the founding General Chair of the IEEE International Conference on Biometrics Theory, Applications and Systems (BTAS), having served as General Chair in 2007, 2008 and 2009. Professor Bowyer earned his Ph.D. in Computer Science from Duke University.

Patrick J. Flynn is Professor of Computer Science & Engineering and Concurrent Professor of Electrical Engineering at the University of Notre Dame. He received the B.S. in Electrical Engineering (1985), the M.S. in Computer Science (1986), and the Ph.D. in Computer Science (1990) from Michigan State University, East Lansing. He has held faculty positions at Notre Dame (1990-1991, 2001-present), Washington State University (1991-1998), and Ohio State University (1998-2001). His research interests include computer vision, biometrics, and image processing. Dr. Flynn is an IEEE Fellow, an IAPR Fellow, and an ACM Distinguished Scientist. He is a past Associate Editor and Associate Editor-in-Chief of IEEE Transactions on PAMI, and a past associate editor IEEE Trans. on Information Forensics and Security, IEEE Trans. on Image Processing, Pattern Recognition, and Pattern Recognition Letters. He has received outstanding teaching awards from Washington State University and the University of Notre Dame.

Soma Biswas is currently working as a Research Assistant Professor at the University of Notre Dame. Her research interests are in signal, image, and video processing, computer vision, and pattern recognition. She received the B.E. degree in electrical engineering from Jadavpur University, Kolkata, India, in 2001, the M.Tech. degree from the Indian Institute of Technology, Kanpur, in 2004, and the Ph.D. degree in electrical and computer engineering from the University of Maryland, College Park, in 2009.