

Face recognition using more than one still image: What is more?

Shaohua Kevin Zhou

Siemens Corporate Research
Integrated Data Systems Department
755 College Road East, Princeton, NJ 08540
Emails: {kzhou}@scr.siemens.com

Abstract. While face recognition from a single still image has been extensively studied over a decade, face recognition based on more than one still image, such as multiple still images or a video sequence, is an emerging topic. Using more than one image introduces new recognition settings. In terms of recognition algorithm, multiple still images or a video sequence can be treated as a single still image in a degenerate manner. However, this treatment neglects additional properties present in multiple still images and/or video sequences. In this paper, we review three properties, manifested in multiple still images and/or video sequence, and their implications to different recognition settings. We also list corresponding approaches proposed in the literature that utilize these properties.

1 Introduction

While face recognition from a single still image has been extensively studied over a decade, face recognition based on a group of still images (also referred as multiple still images) or a video sequence is an emerging topic. This is mainly evidenced by the growing increase in the literature. For instance, a research initiative called Face Recognition Grand Challenge [1] is organized. One specific challenge directly addresses the use of multiple still images that is reportedly to improve the recognition accuracy significantly [3]. Recently a new workshop jointly held with CVPR 2004 is devoted on face processing in video [2]. It is predictable that with the ubiquity of video sequences, face recognition based on video sequences will become more and more popular.

It is obvious that multiple still images or a video sequence can be regarded as a single still image in a degenerate manner. More specifically, suppose that we have a single-still-image-based face recognition algorithm \mathcal{A} (or the base algorithm) by some means, we can construct an assembly recognition algorithm based on multiple still images or a video sequence by combining multiple base algorithms denoted by \mathcal{A}_i 's. Each \mathcal{A}_i takes a different single image y_i as input, coming from the multiple still images or video sequences. The combining rule can be additive, multiplicative, and so on. Section 3 presents a detailed example of constructing an assembly recognition algorithm.

Even though the assembly algorithms might work well in practice, clearly, the overall recognition performance of the assembly algorithm is solely based on those of separate algorithms and hence designing the base algorithm \mathcal{A} is of ultimate importance. Therefore, the assembly algorithms completely neglect additional properties possessed by multiple still images or video sequences.

Three additional properties are available for multiple still images and/or video sequences:

1. [**P1: Multiple observations**]. This property is directly utilized by the assembly algorithm. One main disadvantage of the assembly algorithms is its *ad hoc* combining rule. However, theoretic analysis based on multiple observations can be derived as shown later.
2. [**P2: Temporal continuity/Dynamics**]. Successive frames in the video sequences are continuous in the temporal dimension. Such continuity, coming from facial expression, geometric continuity related to head and/or camera movement, or photometric continuity related to changes in illumination, provides an additional constraint for modeling face appearance. In particular, temporal continuity can be further characterized using dynamics. For example, facial expression and head movement when an individual participates certain activity result in structured changes in face appearance. Depiction of such structured change (or dynamics) further regularizes face recognition.
3. [**P3: 3D model**]. This means that we are able to reconstruct 3D model from a group of still images and a video sequence. Recognition can then be based on the 3D model.

Clearly, the first and third properties are shared by multiple still images and video sequences. The second property is solely possessed by video sequences. We will elaborate these properties in Section 4.

The properties manifested in multiple still images and video sequences present new challenges and opportunities. On the one hand, by judiciously exploiting these features, we can design new recognition algorithms other than those of assembly nature. On the other hand, cares should be exercised when exploiting these properties. Generally speaking, the algorithms utilizing these properties are advantageous to the assembly ones in terms of recognition performance, computational efficiency, etc.

There are two recent survey papers [11, 42] on face recognition in the literature. In [11], face recognition is in its early age and none of the reviewed approaches was video-based. In [42], video-based recognition is identified as one key topic. Even though it had been reviewed quite intensively, all video-based approaches were not categorized. In this paper, we attempt to bring out new insights through studying the three properties. We proceed to the next section by recapitulating some basics of face recognition and introduce new recognition settings that use more than one image.

2 Recognition Setting

In this section, we address the concept of training, gallery, and probe sets and present various recognition settings based on different types of inputs used in the gallery and probe sets.

2.1 Gallery, probe, and training sets

We here follow a face recognition test protocol FERET [28] widely observed in the face recognition literature. FERET assumes availability of the following three sets, namely one training set, one gallery set, and one probe set. The gallery and probe sets are used in the testing stage. The gallery set contains images with known identities and the probe set with unknown identities. The algorithm associates descriptive features with the images in the gallery and probe sets and determines the identities of the probe images by comparing their associated features with those features associated with gallery images.

According to the imagery utilized in the gallery and probe sets, we can define the following nine recognition settings as in Table 1. For instance, the *mStill-to-Video* setting utilizes multiple still images for each individual in the gallery set and a video sequence for each individual in the probe set. The FERET investigated the *sStill-to-sStill* recognition setting.

Probe \ Gallery	A single still image	A group of still images	A video sequence
A single still image	<i>sStill-to-sStill</i>	<i>mStill-to-sStill</i>	<i>Video-to-sStill</i>
A group of still images	<i>sStill-to-mStill</i>	<i>mStill-to-mStill</i>	<i>Video-to-mStill</i>
A video sequence	<i>sStill-to-Video</i>	<i>mStill-to-Video</i>	<i>Video-to-Video</i>

Table 1. Recognition settings based on a single still image, multiple still images and a video sequence.

The need of a training set in addition to the gallery and probe sets is mainly motivated by that fact that the *sStill-to-sStill* recognition setting is used in the FERET. The purpose of the training set is provided for the recognition algorithm to learn the face space. For example, in subspace methods, the training set is used to learn the projection matrix for the face space. Typically, the training set does not overlap with the gallery and probe sets in terms of identity. This is based on that the face space characterization is applied for all individuals and generalization across the identities in the training and gallery sets is used. If more than one still image is used in the gallery set to represent one individual, the training set can be omitted provided that there are enough number of images.

3 Assembly Recognition Algorithm

Here is a concrete example of constructing an assembly recognition algorithm. Suppose that the still-image-based face recognition uses the nearest distance

classification rule, the recognition algorithm \mathcal{A} performs the following:

$$\mathcal{A} : \hat{n} = \arg \min_{n=1,2,\dots,N} d(\mathbf{y}, \mathbf{x}^{[n]}), \quad (1)$$

where N is the number of individuals in the gallery set, d is the distance function, $\mathbf{x}^{[n]}$ represents the n^{th} individual in the gallery set, and \mathbf{y} is the probe single still image. Equivalently, the distance function can be replaced by a similarity function s . The recognition algorithm becomes:

$$\mathcal{A} : \hat{n} = \arg \max_{n=1,2,\dots,N} s(\mathbf{y}, \mathbf{x}^{[n]}). \quad (2)$$

In this paper, we interchange the use of the distance and similarity functions if there is no confusion.

The common choices for d include the following:

- Cosine angle:

$$d(\mathbf{y}, \mathbf{x}^{[n]}) = 1 - \cos(\mathbf{y}, \mathbf{x}^{[n]}) = 1 - \frac{\mathbf{y}^T \mathbf{x}^{[n]}}{\|\mathbf{y}\| \cdot \|\mathbf{x}^{[n]}\|} \quad (3)$$

- Distance in subspace:

$$d(\mathbf{y}, \mathbf{x}^{[n]}) = \|\mathbf{P}^T \{\mathbf{y} - \mathbf{x}^{[n]}\}\|^2 = \{\mathbf{y} - \mathbf{x}^{[n]}\}^T \mathbf{P} \mathbf{P}^T \{\mathbf{y} - \mathbf{x}^{[n]}\}, \quad (4)$$

where \mathbf{P} is a subspace projection matrix. The common subspace methods include principal component analysis (a.k.a. eigenface [37]), linear discriminant analysis (a.k.a. Fisherface [7, 15, 41]), independent component analysis [6], local feature analysis [27], intra-personal subspace [26, 45] etc.

- ‘Generalized’ Mahanalobis distance:

$$d(\mathbf{y}, \mathbf{x}^{[n]}) = \{\mathbf{y} - \mathbf{x}^{[n]}\}^T \mathbf{W} \{\mathbf{y} - \mathbf{x}^{[n]}\}, \quad (5)$$

where the \mathbf{W} matrix plays a weighting role. If $\mathbf{W} = \mathbf{P} \mathbf{P}^T$, then the ‘generalized’ Mahanalobis distance reduces to the distance in subspace. If $\mathbf{W} = \Sigma^{-1}$ with Σ being a covariance matrix, then the ‘generalized’ Mahanalobis distance reduces to the regular Mahanalobis distance.

Using the base algorithm \mathcal{A} defined in (1) as a building block, we can easily construct various assembly recognition algorithms [14] based on a group of still images and a video sequence. By denoting a group of still images and a video sequence by $\{\mathbf{y}_t; t = 1, 2, \dots, T\}$, the recognition algorithm \mathcal{A}_t for \mathbf{y}_t is simply

$$\mathcal{A}_t : \hat{n} = \arg \min_{n=1,2,\dots,N} d(\mathbf{y}_t, \mathbf{x}^{[n]}). \quad (6)$$

Some commonly used combining rules are listed in Table 2.

In the above, the n^{th} individual in the gallery set is represented by a single still image $\mathbf{x}^{[n]}$. This can be generalized to use multiple still images or a video sequence $\{\mathbf{x}_s^{[n]}; s = 1, 2, \dots, K_s\}$. Similarly, the resulting assembly algorithm is to combine the base algorithms denoted by \mathcal{A}_{ts} ’s:

$$\mathcal{A}_{ts} : \hat{n} = \arg \min_{n=1,2,\dots,N} d(\mathbf{y}_t, \mathbf{x}_s^{[n]}). \quad (7)$$

Method	Rule
Minimum arithmetic mean	$\hat{n} = \arg \min_{n=1,2,\dots,N} \frac{1}{T} \sum_{t=1}^T d(\mathbf{y}_t, \mathbf{x}^{[n]})$
Minimum geometric mean	$\hat{n} = \arg \min_{n=1,2,\dots,N} \sqrt[T]{\prod_{t=1}^T d(\mathbf{y}_t, \mathbf{x}^{[n]})}$
Minimum median	$\hat{n} = \arg \min_{n=1,2,\dots,N} \{med_{t=1,2,\dots,T} d(\mathbf{y}_t, \mathbf{x}^{[n]})\}$
Minimum minimum	$\hat{n} = \arg \min_{n=1,2,\dots,N} \{\min_{t=1,2,\dots,T} d(\mathbf{y}_t, \mathbf{x}^{[n]})\}$
Majority voting	$\hat{n} = \arg \max_{n=1,2,\dots,N} \sum_{t=1}^T \mathbf{J}[\mathcal{A}_t(\mathbf{y}) == n]$

Table 2. A list of combining rules. The J function used in majority voting is an indicator function.

4 Properties

The multiple still images and video sequence are different from one still image as they possess additional properties not cherished by a still image. In particular, three properties manifest themselves that motivated various approaches recently proposed in the literature. Below, we analyze the three properties one by one.

4.1 [P1: Multiple observations]

This is the most commonly used feature of multiple still images and video sequence. If only this property is concerned, a video sequence reduces to a group of still images with the temporal dimension stripped. In other words, every video frame is treated as a still image. Another implicit assumption is that all face images are normalized before subjecting to subsequence analysis.

The assembly algorithms utilize this property in a straightforward fashion. However, as mentioned above, the combining rules are rather *ad hoc*, which leaves room for a systematic exploration of this property by using different representations of the multiple observations $\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_T\}$. Once an appropriate representation is fixed, a recognition algorithm can be accordingly designed.

Various ways of summarizing multiple observations have been proposed. In terms of the summarization rule, these approaches can be roughly grouped into four categories.

One image or several images Multiple observations $\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_T\}$ are summarized into one image $\hat{\mathbf{y}}$ or several images $\{\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \dots, \hat{\mathbf{y}}_m\}$ (with $m < T$). For instance, one can use the mean or the median of $\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_T\}$ as the summary image $\hat{\mathbf{y}}$. Clustering techniques can be invoked to produce the summary images $\{\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \dots, \hat{\mathbf{y}}_m\}$. In terms of recognition, we can simply apply the still-image-based face recognition algorithm based on $\hat{\mathbf{y}}$ or $\{\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \dots, \hat{\mathbf{y}}_m\}$. This applies to all nine recognition settings listed in Table 1.

Matrix Multiple observations $\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_T\}$ form a matrix¹ $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_T]$. The main advantages of using the matrix representation is that we can rely on

¹ Here we assume that each image \mathbf{y}_i is ‘vectorized’.

the rich literature of matrix analysis. For example, various matrix decompositions can be invoked to represent the original data more efficiently. Metrics measuring similarity between two matrices can be used for recognition.

This applies to the *mStill-to-mStill*, *Video-to-mStill*, *Video-to-mStill*, and *Video-to-Video* recognition settings. Suppose that the n^{th} individual in the gallery set has a matrix $\mathbf{X}^{[n]}$, we determine the identity of a probe matrix \mathbf{Y} as

$$\hat{n} = \arg \min_{n=1,2,\dots,N} d(\mathbf{Y}, \mathbf{X}^{[n]}), \quad (8)$$

where d is a matrix distance function.

Yamaguchi *et al.* [40] proposed the so-called *Mutual Subspace Method* (MSM) method. In this method, the matrix representation is used and the similarity function between two matrices is defined as the angle between two subspaces of the matrices (also referred as principal angle or canonical correlation coefficient). Wolf and Shashua [38] extended computation of the principal angles into a nonlinear feature space \mathcal{H} called reproducing kernel Hilbert space (RKHS) [32] induced by a positive definite kernel function. Zhou [48] systematically investigated the kernel functions taking matrix as input (also referred to as matrix kernels). Two kernel functions using trace and determinant are proposed. Using them as building blocks, Zhou [48] constructed more kernels based on the column basis matrix, the ‘kernelized’ matrix, and the column basis matrix of the ‘kernelized’ matrix.

Probability density function (PDF) In this rule, multiple observations $\{y_1, y_2, \dots, y_T\}$ are regarded as independent realizations drawn from an underlying distribution. PDF estimation techniques such as parametric, semi-parametric, and non-parametric methods [13] can be utilized.

In the *mStill-to-mStill*, *Video-to-mStill*, *Video-to-mStill*, and *Video-to-Video* recognition settings, recognition can be performed by comparing distances between PDF’s, such as Bhattacharyya and Chernoff distances, Kullback-Leibler divergence, and so on. More specifically, suppose that the n^{th} individual in the gallery set has a pdf $p^{[n]}(\mathbf{x})$, we determine the identity of a probe PDF $q(\mathbf{y})$ as

$$\hat{n} = \arg \min_{n=1,2,\dots,N} d(q(\mathbf{y}), p^{[n]}(\mathbf{x})), \quad (9)$$

where d is a probability distance function.

In the *mStill-to-sStill*, *Video-to-sStill*, *sStill-to-mStill*, and *sStill-to-Video* settings, recognition becomes a hypothesis testing problem. For example, in the *sStill-to-mStill* setting, if we can summarize the multiple still images in query into a pdf, say $q(\mathbf{y})$, then recognition is to test which gallery image $\mathbf{x}^{[n]}$ is mostly likely to be generated by $q(\mathbf{y})$.

$$\hat{n} = \arg \max_{n=1,2,\dots,N} q(\mathbf{x}^{[n]}). \quad (10)$$

Notice that this is different from the *mStill-to-sStill* setting, where each gallery object has a density $p^{[n]}(\mathbf{y})$, then given a probe single still image \mathbf{y} , recognition

checks the following:

$$\hat{n} = \arg \max_{n=1,2,\dots,N} p^{[n]}(\mathbf{y}). \quad (11)$$

Shakhnoarovich *et al.* [33] proposed to use the multivariate normal density for summarizing face appearances and the Kullback-Leibler (KL) divergence or relative entropy for recognition. In [18, 19], Jebara and Kondon proposed probability product kernel function. In [47], Zhou and Chellappa computed the probabilistic distances such as the Chernoff and Bhattacharyya distances, the KL divergence, etc., in the RKHS space. Recently, Arandjelović and Cipolla [5] used resistor-average distance (RAD) for video-based recognition.

Manifold In this rule, face appearances of multiple observations form a highly nonlinear manifold \mathcal{P} . Manifold learning has recently attracted a lot of attention. Examples include [30, 35].

After characterizing the manifolds, face recognition reduces to (i) comparing two manifolds if we are in the *mStill-to-mStill*, *Video-to-mStill*, *Video-to-mStill*, and *Video-to-Video* settings and (ii) comparing distances from one data point to different manifolds if we are in the *mStill-to-sStill*, *Video-to-sStill*, *sStill-to-mStill*, and *sStill-to-Video* settings.

For instance, in the *Video-to-Video* setting, galley videos are summarized into manifolds $\{\mathcal{P}^{[n]}; n = 1, 2, \dots, N\}$. For the probe video that is summarized into a manifold \mathcal{Q} , its identity is determined as

$$\hat{n} = \arg \min_{n=1,2,\dots,N} d(\mathcal{Q}, \mathcal{P}^{[n]}), \quad (12)$$

where d calibrates the distance between two manifolds.

In the *Video-to-sStill* setting, for the probe still image \mathbf{y} , its identity is determined as

$$\hat{n} = \arg \min_{n=1,2,\dots,N} d(\mathbf{y}, \mathcal{P}^{[n]}), \quad (13)$$

where d calibrates the distance from a data point to a manifold.

Fitzgibbon and Zisserman [16] proposed to compute a joint manifold distance to cluster appearances. Li *et al.* [24] proposed identity surface that depicts face appearances presented in multiple poses. A video sequence corresponds to a trajectory traced out in the identity surface. Trajectory matching is used for recognition.

4.2 [**P2: Temporal continuity/Dynamics**]

Property *P1* strips the temporal dimension available in the video sequence. In this property *P2*, we bring back the temporal dimension. Clearly, the property *P2* only holds for video sequence.

Successive frames in the video sequences are continuous in the temporal dimension. The continuity arising from dense temporal sampling is two-fold: the face movement is continuous and the change in appearance is continuous.

Temporal continuity provides an additional constraint for modeling face appearance. For example, smoothness of face movement is used in the face tracking. As mentioned earlier, it is implicitly assumed that all face images are normalized before utilization of the property $P1$ of multiple observations. For the purpose of normalization, face detection is independently applied on each image. When temporal continuity is available, tracking can be applied instead of detection to perform normalization of each video frame.

Temporal continuity also plays an important role for recognition. Recently psychophysical evidence [20] reveals that moving faces are more recognizable. In addition to temporal continuity, face movement and face appearance follow certain dynamics, i.e., changes in movement and appearance are not random. Understanding dynamics is also important for face recognition.

Simultaneous tracking and recognition proposed by Zhou and Chellappa [43, 44] is the first approach that systematically studied how to incorporate temporal continuity in video-based recognition. Zhou and Chellappa modeled two tasks involved, namely tracking and recognition, in one probabilistic framework. They compute the posterior recognition probability $p(n_t, \theta_t | \mathbf{y}_{0:t})$ where n_t is the identity variable, θ_t is the tracking parameter, and $\mathbf{y}_{0:t} = \{y_0, y_2, \dots, y_t\}$ is the video observation. Figure 1 illustrates the recognition results reported in [43, 44].

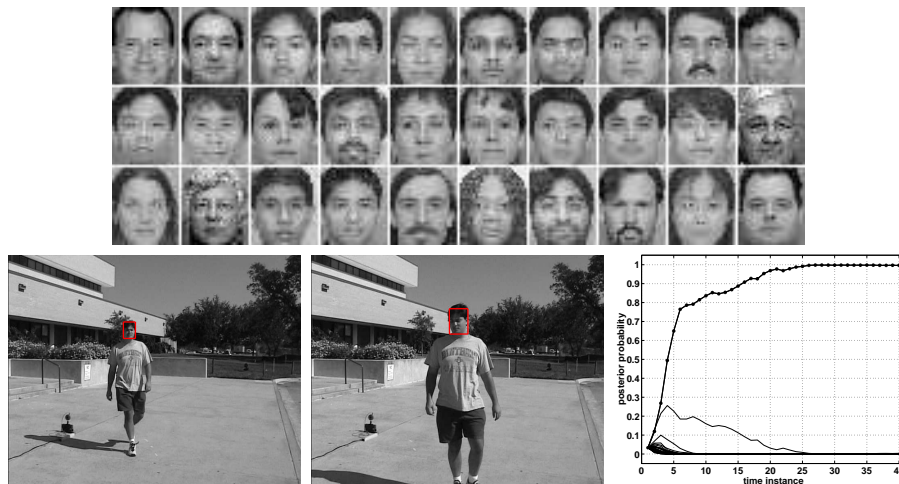


Fig. 1. Top row: the gallery images. Bottom row: The first (left) and the last (middle) frames of the video sequences with tracking results indicated by the box and the posterior probability $p(n_t | \mathbf{y}_{0:t})$.

Krueger and Zhou [21, 44] extended the approach in [43] to handle video sequence in the gallery set. Representative exemplars are learned from the gallery video sequences to depict individuals. Then simultaneous tracking and recognition [43] was invoked to handle video sequences in the probe set. Li and Chellappa

[23] also proposed an approach somewhat similar to [43]. In [23], only tracking was implemented using SIS and recognition scores were subsequently derived based on tracking results.

Lee *et al.* [22] performed video-based face recognition using probabilistic appearance manifolds. The main motivation is to model appearances under pose variation, i.e., a generic appearance manifold consists of several pose manifolds. Liu and Chen [25] proposed to use adaptive HMM to depict the dynamics. Aggarwal *et al.* [4] proposed a system identification approach for video-based face recognition. The face sequence is treated as a first-order auto-regressive and moving averaging (ARMA) random process. Promising experimental results (over 90%) were reported when significant pose and expression variations are present in the video sequences.

Facial expression analysis is also related to temporal continuity/dynamics, but not directly related to face recognition. Approaches to expression analysis include [8, 36].

4.3 [P3: 3D model]

This means that we are able to reconstruct 3D model from a group of still images and a video sequence. This leads to the literature of multiview geometry and structure from motion (SfM). Even though SfM has been studied for a long time, current SfM algorithms are not reliable enough for accurate 3D model reconstruction. Researchers therefore incorporate or solely use prior 3D face models (that are acquired beforehand) to derive the reconstruction result.

The 3D model usually possesses two components: geometric and photometric. The geometric component describes the depth information of the face and the photometric component depicts the texture map.

Recognition can then be performed directly based on the 3D model. More specifically, for any recognition setting, suppose that, galley individuals are summarized into 3D models $\{\mathcal{M}^{[n]}; n = 1, 2, \dots, N\}$. For multiple observations of a probe individual that are summarized into a 3D model \mathcal{N} , its identity is determined as

$$\hat{n} = \arg \min_{n=1,2,\dots,N} d(\mathcal{N}, \mathcal{M}^{[n]}), \quad (14)$$

where d calibrates the distance between two models.

For one probe still image \mathbf{y} , its identity is determined as

$$\hat{n} = \arg \min_{n=1,2,\dots,N} d(\mathbf{y}, \mathcal{M}^{[n]}), \quad (15)$$

where d calibrates the cost of generating a data point from a model.

It is interesting to note that Face Recognition Grand Challenge (FRGC) [1] also proposed the challenge of comparing 3D face models but obtained from a laser scan not from a 3D reconstruction algorithm.

Blanz and Vetter [9] fitted a 3D morphable model to a single still image. The 3D morphable model can be thought of an extension of 2D active appearance model [12] to 3D. However, the 3D morphable model uses a linear combination of

dense 3D models. Xiao *et al.* [39] proposed to integrate a linear combination of 3D sparse model and a 2D appearance model. Examples of SfM for reconstructing the 3D model include [10, 29, 31]. Bundle adjustment [17, 34] is a combination of prior 3D model with SfM.

5 Conclusions

In this paper, we have studied three new properties present in the multiple still images and video sequences. We have also addressed the use of these properties in different recognition settings and briefly reviewed the proposed approaches in the literature.

Studying the recognition algorithms from the perspective of additional properties is very beneficial. In particular, we can forecast new approaches that can be developed to realize the full potentials of multiple still images or video sequences. For example, one can combine these properties [46] to arrive at possibly more accurate algorithms.

References

1. Face Recognition Grand Challenge. <http://bbs.bee-biometrics.org>.
2. The First IEEE Workshop on Face Processing in Video. <http://www.visioninterface.net/fpiv04>.
3. F. Fraser, "Exploring the use of face recognition technology for border control applications - Australia's experience," Biometric Consortium Conference, 2003.
4. G. Aggarwal, A. Roy-Chowdhury, and R. Chellappa, "A system identification approach for video-based face recognition," *Proceedings of International Conference on Pattern Recognition*, Cambridge, UK, August 2004.
5. O. Arandjelović and R. Cipolla, "Face recognition from face motion manifolds using robust kernel resistor-average distance," *IEEE Workshop on Face Processing in Video*, Washington D.C., USA, 2004.
6. M.S. Barlett, H.M. Ladesand, and T.J. Sejnowski, "Independent component representations for face recognition," *Proceedings of SPIE 3299*, pp. 528-539, 1998.
7. P. N. Bellhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, pp. 711-720, 1997.
8. M.J. Black and Y. Yacoob, "Recognizing facial expressions in image sequences using local parameterized models of image motion," *International Journal of Computer Vision*, vol. 25, pp. 23-48, 1997.
9. V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 1063-1074, 2003.
10. M.E. Brand, "Morphable 3D Models from Video," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Hawaii, 2001.
11. R. Chellappa, C. L. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey," *Proceedings of The IEEE*, vol. 83, pp. 705-740, 1995.
12. T.F. Cootes, G.J. Edwards, and C.J. Taylor, "Active appearance models," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681-685, 2001.

13. R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Wiley-Interscience, 2001.
14. G.J. Edwards, C.J. Taylor, and T.F. Taylor, "Improving identification performance by integrating evidence from sequences," *IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, pp. 486-491, Fort Collins, Colorado, USA, 1999.
15. K. Etemad and R. Chellappa, "Discriminant analysis for recognition of human face images," *Journal of Optical Society of America A*, pp. 1724-1733, 1997.
16. A. Fitzgibbon and A. Zisserman, "Joint manifold distance: a new approach to appearance based clustering," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Madison, WI, 2003.
17. P. Fua, "Regularized bundle adjustment to model heads from image sequences without calibrated data," *International Journal of Computer Vision*, vol. 38, pp. 153-157, 2000.
18. T. Jebara and R. Kondor, "Bhattacharyya and Expected Likelihood Kernels," *Conference on Learning Theory, COLT*, 2003.
19. R. Kondor and T. Jebara, "A Kernel Between Sets of Vectors," *International Conference on Machine Learning, ICML*, 2003.
20. B. Knight and A. Johnston, "The role of movement in face recognition," *Visual Cognition*, vol. 4, pp. 265-274, 1997.
21. V. Krueger and S. Zhou, "Exemplar-based face recognition from video," *European Conference on Computer Vision*, Copenhagen, Denmark, 2002.
22. K. Lee, M. Yang, and D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Madison, WI, 2003.
23. B. Li and R. Chellappa, "A generic approach to simultaneous tracking and verification in video," *IEEE Trans. on Image Processing*, vol. 11, no. 5, pp. 530-554, 2002.
24. Y. Li, S. Gong, and H. Liddell, "Constructing face identity surface for recognition," *International Journal of Computer Vision*, vol. 53, no. 1, pp. 71-92, 2003.
25. X. Liu and T. Chen, "Video-based face recognition using adaptive hidden markov models," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Madison, WI, 2003.
26. B. Moghaddam, "Principal manifolds and probabilistic subspaces for visual recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, pp. 780-788, 2002.
27. P. Penev and J. Atick, "Local feature analysis: A general statistical theory for object representation," *Networks: Computations in Neural Systems*, vol. 7, pp. 477-500, 1996.
28. P.J. Phillips, H. Moon, S. Rizvi, and P.J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1090-1104, 2000.
29. G. Qian and R. Chellappa, "Structure from motion using sequential monte carlo methods," *Proceedings of International Conference on Computer Vision*, pp. 614-621, Vancouver, BC, 2001.
30. S.T. Roweis and L.K. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science*, vol. 290, pp. 2323-2326, 2000.
31. A. Roy-Chowdhury and R. Chellappa, "Face reconstruction from video using uncertainty analysis and a generic model," *Computer Vision and Image Understanding*, vol. 91, pp. 188-213, 2003.
32. B. Schölkopf and A. Smola, *Support Vector Learning*. Press, 2002.

33. G. Shakhnarovich, J. Fisher, and T. Darrell, "Face recognition from long-term observations," *European Conference on Computer Vision*, Copenhagen, Denmark, 2002.
34. Y. Shan, Z. Liu, and Z. Zhang "Model-based bundle adjustment with applicaiton to face modeling," *Proceedings of International Conference on Computer Vision*, pp. 645–651, Vancouver, BC, 2001.
35. J.B. Tenenbaum, V. de Silva and J.C. Langford. "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000
36. Y. Tian, T. Kanade, and J. Cohn, "Recognizing action units of facial expression analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, pp. 1-19, 2001.
37. M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience*, vol. 3, pp. 72–86, 1991.
38. L. Wolf and A. Shashua, "Kernel principal angles for classification machines with applications to image sequence interpretation," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Madison, WI, 2003.
39. J. Xiao, S. Baker, I. Matthews, and T. Kanade, "Real-time combined 2D+3D active appearance models," *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington, DC, 2004.
40. O. Yamaguchi, K. Fukui and K. Maeda, "Face recognition using temporal image sequence," *Proceedings of International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, 1998.
41. W. Zhao, R. Chellappa, and A. Krishnaswamy, "Discriminant analysis of principal components for face recognition," *Proceedings of International Conference on Automatic Face and Gesture Recognition*, pp. 361-341, Nara, Japan, 1998.
42. W. Zhao, R. Chellappa, A. Rosenfeld, and P.J. Phillips, "Face recognition: A literature survey," *ACM Computing Surveys*, vol. 12, 2003.
43. S. Zhou and R. Chellappa, "Probabilistic human Recognition from video," *European Conference on Computer Vision*, vol. 3, pp. 681-697, Copenhagen, Denmark, May 2002.
44. 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.
45. S. Zhou, R. Chellappa, and B. Moghaddam "Intra-personal kernel subspace for face recognition ," *Proceedings of International Conference on Automatic Face and Gesture Recognition*, Seoul, Korea, May 2004.
46. S. Zhou and R. Chellappa, "Probabilistic identity characterization for face recognition," *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington D.C., USA, June 2004.
47. S. Zhou and R. Chellappa, "Probabilistic distance measures in reproducing kernel Hilbert space," *SCR Technical Report*, 2004.
48. S. Zhou, "Trace and determinant kernels between matrices," *SCR Technical Report*, 2004.