Face Recognition from Face Motion Manifolds using Robust Kernel Resistor-Average Distance

Ognjen Arandjelović Roberto Cipolla

Department of Engineering
University of Cambridge
Cambridge, UK CB2 1PZ
{oa214,cipolla}@eng.cam.ac.uk

Abstract

In this work we consider face recognition from face motion manifolds. An information-theoretic approach with Resistor-Average Distance (RAD) as a dissimilarity measure between distributions of face images is proposed. We introduce a kernel-based algorithm that retains the simplicity of the closed-form expression for the RAD between two normal distributions, while allowing for modelling of complex, nonlinear manifolds. Additionally, it is shown how errors in the face registration process can be modelled to significantly improve recognition. Recognition performance of our method is experimentally demonstrated and shown to outperform state-of-the-art algorithms. Recognition rates of 97–100% are consistently achieved on databases of 35–90 people.

1. Introduction

Important practical applications of automatic face recognition have made it a very popular research area in the last three decades, see [1, 4, 7, 22] for surveys. Most of the methods developed deal with *single-shot* recognition. In controlled imaging conditions (lighting, pose and/or occlusions) many have demonstrated good (nearly perfect) recognition results [22]. On the other hand, single-shot face recognition in uncontrolled, or loosely controlled conditions still poses a significant challenge [22].

The nature of many practical applications is such that more than a single image of a face is available. In surveillance, for example, the face can be tracked to provide a temporal sequence of a moving face. In access control use of face recognition the user may be assumed to be cooperative and hence can be instructed to move in front of a fixed camera. Regardless of the setup in which multiple images of a face are acquired, it is clear that this abundance of information can be used to achieve greater robustness of face recognition by resolving some of the inherent ambiguities of the single-shot recognition problem.

The organization of this paper is as follows. Section 2 reviews the existing literature on face recognition from video. Section 3 introduces the concept of classification using the Kernel RAD. In Section 4 we show how errors in the face registration process can be modelled and incorporated in the described recognition framework. Section 5 reports experimental results and compares the proposed method with several competing methods reported in the literature. Finally, Section 6 concludes the paper and discusses promising directions for future research.

2. Related Previous Work

Single-shot face recognition is a well established research area. Algorithms such as Bayesian Eigenfaces [10, 14], Fisherfaces [20, 22], Elastic Bunch Graph Matching [3, 12] or the 3D Morphable Model [2, 15] have demonstrated good recognition results when illumination and pose variations are not large. However, all existing single-shot methods suffer from the limited ability to generalize to unseen illumination conditions or pose.

Compared to single-shot recognition, face recognition from video is a relatively new area of research. Most of the existing algorithms perform recognition from image sequences, using the temporal component to enforce prior knowledge on likely head movements. In the algorithm of of Zhou *et al.* [23] the joint probability distribution of identity and motion is modelled using sequential importance sampling, yielding the recognition decision by marginalization. In [13] Lee *et al.* approximate face manifolds by a finite number of infinite extent subspaces and use temporal information to robustly estimate the operating part of the manifold.

There are fewer methods that recognize from manifolds without the associated ordering of face images, which is the problem we address in this paper. Two algorithms worth mentioning are the Mutual Subspace Method (MSM) of Ya-

maguchi *et al.* [9, 19] and the Kullback-Leibler divergence based method of Shakhnarovich *et al.* [17].

In MSM, infinite extent linear subspaces are used to compactly characterize face sets i.e. the manifolds that they lie on. Two sets are then compared by computing the first three principal angles between corresponding principal component analysis (PCA) subspaces [9]. Varying recognition results were reported using MSM, see [9, 17, 18, 19]. The major limitation of MSM is its simplistic modelling of manifolds of face variation. Their high nonlinearity (see Figure 1) invalidates the assumption that data is well described by a linear subspace. More subtly, the nonlinearity of modelled manifolds means that the PCA subspace estimate is very sensitive to the particular choice of training samples. For example, in the original paper [19] in which face motion videos were used, the estimates are sensitive to the extent of rotation in a particular direction. Finally, MSM does not have a meaningful probabilistic interpretation.

The Kullback-Leibler divergence (KLD) based method [17] is founded on information-theoretic grounds. In the proposed framework, it is assumed that i-th person's face patterns are distributed according to $p_i(\mathbf{x})$. Recognition is then performed by finding $p_j(\mathbf{x})$ that best explains the set of input samples – quantified by the Kullback-Leibler divergence. The key assumption in their work, that makes divergence computation tractable, is that face patterns are normally distributed i.e. $p_i(\mathbf{x}) = \mathcal{N}(\bar{\mathbf{x}}_i, \mathbf{C}_i)$. This is a crude assumption (see Figure 1), which explains the somewhat poor results reported with this method [18]. KLD was also criticized for being asymmetric [11].

3. Recognition using Kernel RAD

3.1. Resistor-Average Distance

Resistor-Average Distance is a symmetric measure of dissimilarity of two probability distributions. It has a close relationship to optimal classifier performance and reflects its error rate better than the KLD from which it is derived [11]. It is defined as:

$$D_{RAD}(p,q) = \left(D_{KL}(p||q)^{-1} + D_{KL}(q||p)^{-1}\right)^{-1} \quad (1)$$

KLD is an information-theoretic measure that quantifies how well a particular pdf $p(\mathbf{x})$ describes samples from another pdf $q(\mathbf{x})$ [5]. It is defined as:

$$D_{KL}(p||q) = \int p(\mathbf{x}) \log_2 \left(\frac{p(\mathbf{x})}{q(\mathbf{x})}\right) d\mathbf{x}. \quad (2)$$

For most practical purposes the evaluation of the above expression is computationally expensive and numerically problematic. However, when $p(\mathbf{x})$ and $q(\mathbf{x})$ are two normal distributions, there is a closed-form expression for KLD [21]:

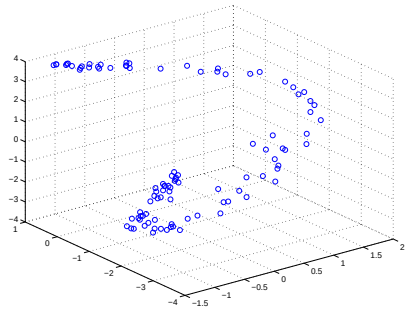

Figure 1: A typical face manifold of lateral head rotation around the fronto-parallel face ($\pm 30°$). Shown is a projection to the first 3 principal components. The manifold can be seen to be smooth, but highly nonlinear.

$$D_{KL}(p||q) = \frac{1}{2}\log_2\left(\frac{|\mathbf{\Sigma}_q|}{|\mathbf{\Sigma}_p|}\right) + \frac{1}{2}\mathrm{Tr}\left(\mathbf{\Sigma}_p\mathbf{\Sigma}_q^{-1} + \mathbf{\Sigma}_q^{-1}(\bar{\mathbf{x}}_q - \bar{\mathbf{x}}_p)(\bar{\mathbf{x}}_q - \bar{\mathbf{x}}_p)^T\right) - \frac{N}{2} \quad (3)$$

where N is the dimensionality of data, $\bar{\mathbf{x}}_p$ and $\bar{\mathbf{x}}_q$ data means, and $\mathbf{\Sigma}_p$ and $\mathbf{\Sigma}_q$ the corresponding covariance matrices.

The simplicity of (3) comes at the cost of a strong assumption on the two data distributions. In the case of variation of face patterns (vectors of pixel values), this assumption is unjustified. Lighting or pose changes, or even simple plane transformations (rotation, translation), are all highly nonlinear (see Figure 1). In the next section we review Kernel PCA (KPCA) which is in our method used to efficiently handle nonlinear face manifolds.

3.2. Kernel Principal Component Analysis

PCA is a technique in which an orthogonal basis transformation is applied such that the data covariance matrix $\mathbf{C} = \langle(\mathbf{x}_i - \langle\mathbf{x}_j\rangle)(\mathbf{x}_i - \langle\mathbf{x}_j\rangle)^T\rangle$ is diagonalized. In the case of nonlinearly distributed data, PCA does not capture the true modes of variation well.

The idea behind KPCA is to map data to a high-dimensional space in which it *is* approximately linear – then the true modes of data variation can be found using standard PCA. Performing this mapping explicitly is prohibitive for computational reasons. This is why a technique known as the "kernel trick" is used. Let $\mathbf{\Phi}$ map the original data in input space to a high-dimensional pattern space in which it is (approximately) linear. In KPCA the mapping Φ is restricted to be such that there is a function k (the kernel) such that $\mathbf{\Phi}(\mathbf{x}_i)^T\mathbf{\Phi}(\mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j)$. In this case, principal components of the data can be found by performing computations in input space only.

Assuming zero-centred data in the feature space (for information on centring data in the feature space as well as a more detailed treatment of KPCA see [16]), the problem of finding principal components in the feature space is equivalent to solving the eigenvalue problem:

$$\mathbf{K}\mathbf{u}_i = \lambda_i \mathbf{u}_i \tag{4}$$

where $\mathbf{K}$ is the kernel matrix:

$$\mathbf{K}_{j,k} = k(\mathbf{x}_j, \mathbf{x}_k) = \mathbf{\Phi}(\mathbf{x}_j)^T \mathbf{\Phi}(\mathbf{x}_k) \tag{5}$$

The projection of a data point $\mathbf{x}$ to the i-th kernel principal component is computed using the following expression [16]:

$$a_i = \sum_{m=1}^{N} u_i^{(m)} k(\mathbf{x}_m, \mathbf{x}) \tag{6}$$

3.3. Combining RAD and KPCA

The variation of face patterns is highly nonlinear (see Figure 1). Hence RAD between two sparsely sampled face manifolds cannot be easily computed in the input space. Therefore, we are looking for a mapping of data from the input space into a space in which data is nearly linear. As before, we would not like to compute this mapping explicitly. Also, data covariance matrices and their determinants in the expression for the KLD between two normal distributions (3) limit the maximal practical dimensionality of the pattern space.

In our method both of these problems are solved using KPCA. The key observation is that regardless of how high the pattern space dimensionality is, the data has covariance in at most N directions, where N is the number of data points.

Therefore, given two data sets of faces, each describing a smooth manifold, we first find the kernel principal components of their union. After dimensionality reduction is performed by projecting the data onto the first M kernel principal components, the RAD between the two distributions, each assumed Gaussian, is computed.

4. Modelling Registration Errors

The variation of face patterns is highly nonlinear even for simple planar transformations, like translation or rotation. Since reliable methods for facial feature localization have been developed [8], these unwanted variations are best dealt with directly, that is, by recovering transformation parameters from sets of point correspondences. Images can then be registered to have relevant facial features in selected canonical locations.

Most face recognition methods do not consider how registration errors impact recognition performance. Modelling of even small affine misregistrations (by randomly perturbing manually registered data) shows that these variations can be significant, suggesting that they should be included in the data formation model.

We recognize two sources of registration errors:

- errors due to small localization inaccuracies of facial feature detectors, and
- large errors due to wrongly localized features (outliers).

The fundamentally different nature of these two sources of error suggest separate modelling of the two. We handle small registration errors by augmenting the input data sets with synthetically perturbed data. Large registration errors are handled using RANSAC for KPCA computation.

4.1. Small Registration Errors

Small registration errors occur due to small localization inaccuracies of facial feature detectors (see Figure 3). These are most pronounced when input images have low resolution, which is sometimes all that is available, or when subsampling is necessary for computational efficiency reasons. Although at first sight small, these localization errors cause a significant drift of face patterns along the misregistration manifold, especially when few facial features are used or when data is nonlinearly mapped, as in the proposed algorithm.

In our method, a set of face images is augmented by synthetic samples from the corresponding misregistration manifold. The samples are generated by applying small, random perturbations to the input images. We use $N_s \sim 2^{N_p}$ as a rule of thumb for the number of synthetic samples, where N_p is the number of registration transformation parameters (so $N_p = 6$ for affine registration).

4.2. Large Registration Errors

Large registration errors are due to wrong facial feature localization. Unlike small errors from the previous section, these do not lie on a smooth manifold and are in our method considered outliers (see Figure 3).

Our algorithm uses RANSAC [6] for robust estimation of the KPCA subspace used for dimensionality reduction (Section 3.3). This process is summarized in Algorithm 1.

A summary of the complete method proposed is given in Algorithm 2.

5. Experimental Evaluation

We performed several experiments for the purpose of evaluating our algorithm and comparing its performance with

Algorithm 1 RANSAC Kernel PCA

1: $c = 0$
2: **for** $c = 0$ to C_{limit} **do**
3: Randomly select N_d data samples $\{y_i\}$, where N_d is the dimensionality of the KPCA subspace being constructed.
4: Compute kernel principal components of $\{y_i\}$.
5: From the rest of the data select data points within a threshold distance from the origin in the KPCA space. If the number of selected points is greatest so far, call it $\{z_i\}$.
6: **end for**
7: Compute KPCA on $\{z_i\}$.

Algorithm 2 Robust Kernel RAD

Input: sets $\{\mathbf{a}_i\}$, $\{\mathbf{b}_i\}$
Output: $D_{RAD}(\{\mathbf{a}_i\}, \{\mathbf{b}_i\})$

1: Using RANSAC find inliers in $\{\mathbf{a}_i\}$, $\{\mathbf{b}_i\}$ $\rightarrow$ inliers $\{\mathbf{a}_i^V\}$, $\{\mathbf{b}_i^V\}$
2: Perturb $\langle \mathbf{a^V} \rangle$, $\langle \mathbf{b^V} \rangle$ $\rightarrow$ synthetic data $\{\mathbf{a_i^S}\}$, $\{\mathbf{b_i^S}\}$
3: Perform RANSAC Kernel PCA on $\{\mathbf{a}_i^V\} \cup \{\mathbf{b}_i^V\} \cup \{\mathbf{a}_i^S\} \cup \{\mathbf{b}_i^S\}$ $\rightarrow$ principal components $\mathbf{u}_i$
4: Project $\{\mathbf{a}_i^V\} \cup \{\mathbf{a}_i^S\}$, $\{\mathbf{b}_i^V\} \cup \{\mathbf{b}_i^S\}$ onto $\mathbf{u}_i \rightarrow$ projections $\{\mathbf{a}_i^P\}$, $\{\mathbf{b}_i^P\}$
5: Compute RAD between $\{\mathbf{a}_i^P\}$ and $\{\mathbf{b}_i^P\}$ $\rightarrow$ $D_{RAD}(\{\mathbf{a}_i\}, \{\mathbf{b}_i\})$

algorithms in the literature. The data sets used in experiments are described in Section 5.1. Algorithms chosen for comparison are:

- Kernel RAD,
- Robust Kernel RAD,
- KLD-based algorithm of Shakhnarovich *et al.* [17],
- Kernelized KLD-based algorithm,
- Mutual Subspace Method [19],
- Majority vote using Eigenfaces.

The dimensionality of the KPCA subspace used in our Kernel RAD methods was set to 20 for computational reasons. The RBF kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-0.6(\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_i - \mathbf{x}_j))$ was used in all kernel methods and was found empirically. In the original KLD-based method 85% of data energy was explained by the principal subspace used. In MSM, the dimensionality of PCA subspaces was set to 9 [19]. The first 3 principal angles were used for recognition, as this produced best results in the literature [19]. In the Eigenfaces method, the 22-dimensional principal subspace used explained 90% of data energy.

Figure 2: Frames from typical input video sequences used for evaluation of methods in this paper. The bottom-right frame is shown with automatically detected pupils and nostrils, and the region of the face used for recognition.

5.1. Data

The evaluation of methods in this paper was done on six databases, five with 35 and one with 90, mostly male individuals. For each person in a database we collected a training and a testing image set, each consisting of 30-50 images of a face in random motion (yaw within approximately $\pm 30°$), sampled from a video at 10fps (see Figure 2). Illumination conditions for each database were different, but unchanging across the training and testing sets. Face images were affine registered using 4 point correspondences (using pupils and nostrils) and automatically cropped at approximately mouth and mid-forehead level. Pupils and nostrils were automatically detected using the algorithm described in [8]. Finally, for computational and memory reasons, images were subsampled to 15×15 pixel grayscale images with pixel values normalized to lie in the range $[0, 1]$. See Figures 2 and 3. We emphasize that the whole process is automatic – no human intervention is required at any point.

5.2. Results

The recognition results are summarized in Figure 4. Our Robust Kernel RAD outperformed other methods on each database, producing the highest average recognition score

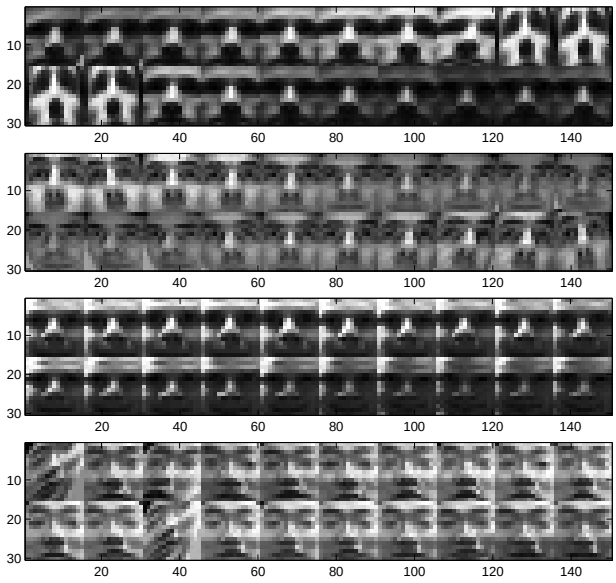

Figure 3: Registered and automatically cropped faces (15×15 pixels) from 4 typical sequences used for comparison of recognition methods in this paper. Presence of outliers can be seen in the last set of faces, while small registration errors are present in all 4 sets.

of 98%, with non-decayed performance for the large database. The original KLD-based method produced very low recognition rates. The likely reason for this is the high nonlinearity of manifolds described by training sets used, caused by close, office lighting.

It can be seen that kernelization alone of the original KLD method improved the results dramatically – from 45% to 85%. This confirms that face recognition by manifold modelling is indeed a promising direction of research. Consistent improvement of recognition rates using RAD over KLD is also demonstrated.

Registration error modelling was confirmed to be very important, additionally decreasing the error rate to 2%. Distance matrices on a typical database are compared for the Kernel RAD method without and with misregistration modelling in Figure 5. Increase not only in recognition performance can be seen, but also in separability of within and between class distances. Separability for the Robust Kernel RAD method is shown in Figure 6.

6. Summary and Conclusions

In this paper we introduced a novel approach to face recognition from face motion manifolds. In the proposed algorithm the Resistor-Average distance computed on nonlinearly mapped data using Kernel PCA is used as a dissimilarity measure between distributions of face images. Additionally, sources of registration errors are explicitly modelled

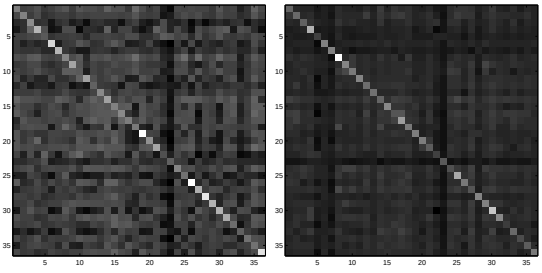

Figure 5: Typical distance matrices using Kernel Relative RAD without (left) and with(right) modelling of registration errors for a set of 35 people. Brighter pixels represent smaller distances (the intensity scale is logarithmic).

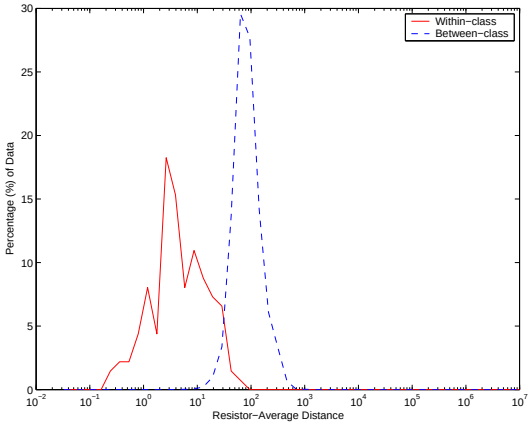

Figure 6: Histograms of within-class (solid line) and between-class (dashed line) Robust Kernel RAD on the 5 databases used in the method evaluation. Good separability is demonstrated.

allowing for their effects to be greatly reduced. A number of experiments was presented using frames from videos of faces in random motion. It was shown that our method consistently outperformed existing methods, achieving on average the recognition rate of 97–100% on 6 databases of 35–90 people each.

The success of recognition based on kernelizing the RAD suggests that future research should concentrate on better understanding of face manifolds. Insight into the shape of these manifolds should help in the choice of more appropriate kernel functions. At last, the proposed method does not handle illumination changes. Our future work will concentrate on efficient illumination compensation on face manifolds.

	Database 1	Database 2	Database 3	Database 4	Database 5	Database 6	Average
Robust Kernel RAD	100	94	97	100	100	97	98
MSM	100	88	94	94	97	94	94
Kernel RAD	90	85	92	92	97	88	91
Kernel KLD	84	85	86	81	88	88	85
Majority vote, Eigenfaces	81	70	58	69	73	73	71
KLD	42	35	39	50	60	44	45

Figure 4: Results of the comparison of our novel algorithm with existing methods in the literature. Shown is the identification rate (%) for five databases of 35 people (1–5) and one of 90 (6). Robust Kernel RAD produced best results on all 6 databases.

Acknowledgements

We would like to thank the Toshiba Corporation for their kind support for our research and the people from the Cambridge University Engineering Department who volunteered to have their face videos entered in our face database. Our thanks also go to Phil Torr, discussions with whom greatly helped us in this research.

References

[1] W. A. Barrett. A survey of face recognition algorithms and testing results. *Systems and Computers*, 1:301–305, 1998.

[2] V. Blanz and T. Vetter. Face recognition based on fitting a 3D morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1063–1074, 2003.

[3] D. S. Bolme. Elastic bunch graph matching. Master's thesis, Colorado State University, 2003.

[4] R. Chellappa, C. L. Wilson, and S. Sirohey. Human and machine recognition of faces: A survey. *Proceedings of the IEEE*, 83(5):705–740, 1995.

[5] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley, 1991.

[6] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *IEEE Transactions on Computers*, 24(6):381–395, 1981.

[7] T. Fromherz, P. Stucki, and M. Bichsel. A survey of face recognition. *MML Technical Report.*, (97.01), 1997.

[8] K. Fukui and O. Yamaguchi. Facial feature point extraction method based on combination of shape extraction and pattern matching. *Systems and Computers in Japan*, 29(6):2170–2177, 1998.

[9] K. Fukui and O. Yamaguchi. Face recognition using multi-viewpoint patterns for robot vision. *Int'l Symp. of Robotics Research*, 2003.

[10] R. Gross, J. She, and J. F. Cohn. Quo vadis face recognition. *Workshop on Empirical Evaluation Methods in Computer Vision*, 1:119–132, 2001.

[11] D. H. Johnson and S. Sinanović. Symmetrizing the Kullback-Leibler distance. *Technical report, Rice University*, 2001.

[12] B. Kepenekci. *Face Recognition Using Gabor Wavelet Transform.* PhD thesis, The Middle East Technical University, 2001.

[13] K. Lee, M. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 313–320, 2003.

[14] B. Moghaddam, W. Wahid, and A. Pentland. Beyond eigenfaces - probabilistic matching for face recognition. *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 30–35, 1998.

[15] S. Romdhani, V. Blanz, and T. Vetter. Face identification by fitting a 3D morphable model using linear shape and texture error functions. *In Proc. IEEE European Conference on Computer Vision*, pages 3–19, 2002.

[16] B. Schölkopf, A. Smola, and K. Müller. Kernel principal component analysis. *Advances in Kernel Methods - SV Learning*, pages 327–352, 1999.

[17] G. Shakhnarovich, J. W. Fisher, and T. Darrel. Face recognition from long-term observations. *In Proc. IEEE European Conference on Computer Vision*, pages 851–868, 2002.

[18] L. Wolf and A. Shashua. Learning over sets using kernel principal angles. *Journal of Machine Learning Research*, 4(10):913–931, 2003.

[19] O. Yamaguchi, K. Fukui, and K. Maeda. Face recognition using temporal image sequence. *IEEE International Conference on Automatic Face and Gesture Recognition*, (10):318–323, 1998.

[20] W. S. Yambor. Analysis of PCA-based and fisher discriminant-based image recognition algorithms. Master's thesis, Colorado State University, 2000.

[21] S. Yoshizawa and K. Tanabe. Dual differential geometry associated with Kullback-Leibler information on the gaussian distributions and its 2-parameter deformations. *SUT Journal of Mathematics*, 35(1):113–137, 1999.

[22] W. Zhao, R. Chellappa, A. Rosenfeld, and P. J. Phillips. Face recognition: A literature survey. *UMD CFAR Tech. Report CAR-TR-948*, 2000.

[23] S. Zhou, V. Krueger, and R. Chellappa. Probabilistic recognition of human faces from video. *Computer Vision and Image Understanding*, 91(1):214–245, 2003.