Face Sequence Recognition Using Grassmann Distances and Grassmann Kernels

Ryosuke Shigenaka, Bisser Raytchev, Toru Tamaki and Kazufumi Kaneda
Department of Information Engineering, Graduate School of Engineering,
Hiroshima University, Hiroshima, JAPAN 739-8527
Email: shigenaka@eml.hiroshima-u.ac.jp, {bisser, tamaki, kin}@hiroshima-u.ac.jp

***Abstract*—In this paper we show how Grassmann distances and Grassmann kernels can be efficiently used to learn and classify face sequence videos. We propose two new methods, the Grassmann Distance Mutual Subspace Method (GD-MSM) which uses Grassmann distances to define the similarity between subspaces of images, and the Grassmann Kernel Support Vector Machine (GK-SVM), which applies two Grassmann kernels – the projection kernel and the Binet-Cauchy kernel – in a convex optimization scheme, using the Support Vector Machine (SVM) framework. GD-MSM and GK-SVM are compared in a face recognition task with several related methods using a large database of face image sequences from 100 subjects, containing expression changes related to a natural conversation setting. Additionally, we study the effect of combining all available training image sequences into a single subspace per category, in comparison with using multiple smaller subspaces, i.e. representing each category by several different subspaces, where each subspace is formed from image sequences taken under different conditions.**

***Index Terms*—Face Sequence Recognition, Grassmann Manifold, Grassmann Distance, Grassmann Kernel, Canonical Angles, Canonical Correlations, Subspace Methods**

I. Introduction

In many computer vision applications, the objects of interest are naturally represented as sets of images, where each image set captures some of the variation of the object under one or several external factors like changes in view angle, illumination conditions, etc., or due to rigid motion or non-rigid deformation of the object itself. Especially when object recognition is concerned, it is generally accepted that it is advantageous to consider the relations between whole image sets, rather than between individual images. To be able to compare two image sets, a suitable distance, or similarity measure, has to be defined. For this purpose, both parametric model-based approaches, and non-parametric sample-base methods have been previously proposed in the literature. In the parametric modeling approach [1], [2], image sets are represented by parametric distribution function, and distances between sets are measured by the Kullback-Leibler (KL) divergence. In the non-parametric sample-based methods [3], [4], typically the nearest-neighbor (NN) distance between the individual samples from each set, or variants of the Hausdorff distance are used.

Recently, an alternative approach is gaining popularity, where image sets are approximated as low-dimensional linear subspaces, and the distance between a pair of image sets is represented as the distance between their corresponding subspaces, a concept which has been well studied before [5], [6]. The distance between subspaces can be represented in terms of the principal angles between them (details are reviewed in the next section). Several methods have been proposed in the literature which make use of the principal angles between subspaces. In the Mutual Subspace Method (MSM) [7], the cosine of the smallest principal angle is used to determine the similarity between two image sequences. Further modifications of MSM include methods like Kernel MSM (KMSM) [8], Constrained MSM (CMSM) [9], [10], Boosted Manifold Principal Angles (BoMPA) [11], Discriminant analysis of Canonical Correlations (DCC) [12], etc.

A recent work in [13] attempts to provide a unifying view on subspace-based learning methods, by formulating the problem on the Grassmann manifold [14], the set of fixed-dimensional linear subspaces of a Euclidean space. Various distances which consider the geometric structure of the Grassmann manifold and can be represented in terms of the principal angles between subspaces have been given in [15] and reviewed in [13]. In [13] it is also shown that the projection metric and the Binet-Cauchy metric can be considered valid metrics on Grassmann manifolds, and kernel functions compatible with these metrics are defined, the projection kernel and the Binet-Cauchy kernel. Then these Grassmann kernels are used in a kernel LDA (Linear Discriminant Analysis) framework. The resulting method is called Grassmann Discriminant Analysis (GDA). Kernel Grassmannian Discriminant Analysis (KGDA) has also been proposed in [16], to extend GDA in a similar manner as KMSM extends MSM by using the kernel trick.

In this paper, we propose to extend the Mutual Subspace Method, so that rather than considering only the cosine of the smallest principal angle, which geometrically is not a good measurement for subspace similarity approximation, and also is not optimal for recognition, the Grassmann distances are used to define the similarity between image sequences. The resulting method we call Grassmann Distance Mutual Subspace Method (GD-MSM). Additionally, in order to obtain more discriminative learning function based on Grassmann distances, we apply the Grassmann kernels from [13] in a convex optimization scheme, using the Support Vector Machine (SVM) framework [17]. The resulting method we call Grassmann Kernel Support Vector Machine (GK-SVM). We compare the performance of GD-MSM and GK-SVM

with several related methods on a large database of facial image sequences. This dataset contains face sequences from 100 subjects with expression changes related to a natural conversation setting (the MOBIO database [18]). Also, we study the effect of combining all available image sequences into one large dictionary (learning a common subspace from all available image sequences for each subject), in comparison with using multiple smaller subspaces, i.e. representing each subject by a multitude of different subspaces, where each subspace is formed from image sequences taken under different conditions.

II. Preliminaries

In this section we provide a brief review of the terminology related to Grassmann distances and Grassmann kernels.

Consider two image sequences $\boldsymbol{S}_i = \{\boldsymbol{x}_1^i, \boldsymbol{x}_2^i, \cdots, \boldsymbol{x}_n^i\}$ and $\boldsymbol{S}_j = \{\boldsymbol{x}_1^j, \boldsymbol{x}_2^j, \cdots, \boldsymbol{x}_m^j\}$, where $\boldsymbol{x}_n^i$ is the *n*-th image in the *i*-th sequence (or an object of interest, e.g. a face, extracted from this image), in vector form. We can represent the whole set of images (or objects) in each sequence by their corresponding subspaces in the Euclidean space $\mathbb{R}^D$, $span(Y_i) = span\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_p\}$ and $span(Y_j) = span\{\boldsymbol{v}_1, \cdots, \boldsymbol{v}_q\}$, where $span(Y_i)$ denotes the subspace spanned by the column vectors of the $D \times p$ matrix $\boldsymbol{Y}_i = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_p]$ and $\{\boldsymbol{u}_1, \cdots, \boldsymbol{u}_p\}$ are orthonormal (they can be computed as the eigenvectors corresponding to the largest p eigenvalues of the correlation matrix obtained from all images in the relevant image sequence). The set of all m-dimensional linear subspaces in $\mathbb{R}^D$ is called the Grassmann manifold $\mathcal{G}(m, D)$, and the subspaces $span(Y_i)$ and $span(Y_j)$ can be considered as two points on the manifold $\mathcal{G}(m, D)$. Various distances on $\mathcal{G}(m, D)$ have been defined [15], all of which can be represented in terms of the principal angles between subspaces [6].

The *principal angles*, or *canonical angles*, $0 \leq \theta_1 \leq \ldots \leq \theta_m \leq \pi/2$, between the subspaces $span(Y_1)$ and $span(Y_2)$ can be defined recursively as

$$\cos\theta_k = \max_{\boldsymbol{u}_k \in span(\boldsymbol{Y}_1)} \max_{\boldsymbol{v}_k \in span(\boldsymbol{Y}_2)} \boldsymbol{u}_k^T \boldsymbol{v}_k, \tag{1}$$

subject to the constraints

$$\begin{aligned} &\boldsymbol{u}_k^T \boldsymbol{u}_k = 1, \boldsymbol{v}_k^T \boldsymbol{v}_k = 1, \\ &\boldsymbol{u}_k^T \boldsymbol{u}_i = 0, \boldsymbol{v}_k^T \boldsymbol{v}_i = 0, (i = 1, \cdots, k-1) \end{aligned} \tag{2}$$

The principal angles can be computed in a numerically stable way from the Singular Value Decomposition (SVD) of $\boldsymbol{Y}_1^T \boldsymbol{Y}_2$

$$\boldsymbol{Y}_1^T \boldsymbol{Y}_2 = \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{V}^T, \boldsymbol{\Lambda} = diag(\lambda_1, \cdots, \lambda_m) \tag{3}$$

where the orthonormal matrices $\boldsymbol{Y}_1$ and $\boldsymbol{Y}_2$ are the matrix representation of $span(Y_1)$ and $span(Y_2)$, and $\lambda_i = \cos\theta_i$ are the cosines of the principal angles θ_i, also known as the *canonical correlations*.

In the rest of the paper we will consider the following Grassmann distances (we follow the notation introduced in [13]), which provide a similarity measure between two subspaces. All of the distances can be represented in terms of the principal angles between the subspaces.

1) The projection metric (the 2-norm of the sines of the principal angles)

$$d_P(Y_1, Y_2) = (\sum_{i=1}^{m} \sin^2\theta_i)^{\frac{1}{2}} = (m - \sum_{i=1}^{m} \cos^2\theta_i)^{\frac{1}{2}} \tag{4}$$

2) The Binet-Cauchy metric (using the product of the cosines of the principal angles)

$$d_{BC}(Y_1, Y_2) = (1 - \prod_i \cos^2\theta_i)^{\frac{1}{2}} \tag{5}$$

3) The Max Correlation (uses only the smallest principal angle, equivalent to the MSM method [7]

$$d_{Max}(Y_1, Y_2) = (1 - \cos^2\theta_1)^{\frac{1}{2}} = \sin\theta_1 \tag{6}$$

4) The Min Correlation (uses only the sine of the largest principal angle)

$$d_{Min}(Y_1, Y_2) = (1 - \cos^2\theta_m)^{\frac{1}{2}} = \sin\theta_m \tag{7}$$

5) The Procrustes (chordal) distance (the minimum distance between different representations of two subspaces $span(Y_1)$ and $span(Y_2)$, using the Frobenius norm)

$$\begin{aligned} d_{CF}(Y_1, Y_2) &= \min_{R_1, R_2 \in O(m)} ||Y_1 R_1 - Y_2 R_2||_F \\ &= 2(\sum_{i=1}^{m} \sin^2(\theta_i/2))^{\frac{1}{2}} \end{aligned} \tag{8}$$

6) The Procrustes (chordal) distance using the matrix 2-norm

$$\begin{aligned} d_{C2}(Y_1, Y_2) &= \min_{R_1, R_2 \in O(m)} ||Y_1 R_1 - Y_2 R_2||_2 \\ &= 2\sin(\theta_m/2) \end{aligned} \tag{9}$$

7) The geodesic distance (the length of the shortest geodesic connecting two points on the Grassmann manifold)

$$d_G(Y_1, Y_2) = \sum_{i=1}^{m} \theta_i^2 \tag{10}$$

8) The mean distance (this is not considered in [13])

$$d_{Mean}(Y_1, Y_2) = \frac{1}{m} \sum_{i=1}^{m} \sin^2\theta_i \tag{11}$$

The distances above can be used to extend the Mutual Subspace Method, by replacing the max-correlation distance (that is, (6) above, which considers only the largest principal angle) with any of the other distances. In section 4 we provide an experimental comparison between these eight Grassmann distances on a face recognition problem using face image sequences.

In [13] it is shown that the projection metric (4) and the Binet-Cauchy metric (5) can be used to define the following positive definite Grassmann kernels:

1) The projection kernel

$$k_P(Y_1, Y_2) = trace[(\boldsymbol{Y}_1\boldsymbol{Y}_1^T)(\boldsymbol{Y}_2\boldsymbol{Y}_2^T)] = ||\boldsymbol{Y}_1^T\boldsymbol{Y}_2||_F^2 \quad (12)$$

2) The Binet-Cauchy kernel

$$k_{BC}(Y_1, Y_2) = \det(\boldsymbol{Y}_1^T\boldsymbol{Y}_2)^2 = \det(\boldsymbol{Y}_1^T\boldsymbol{Y}_2\boldsymbol{Y}_2^T\boldsymbol{Y}_1) = \prod_i \cos^2\theta_i \quad (13)$$

These kernels can be used in conjunction with any of the available kernel-based algorithms [19], and in [13] are used with kernel LDA. In the next section we propose to use them in a convex optimization scheme, in the Support Vector Machine (SVM) framework [20].

III. Methods

Here we describe the methods whose performance on two different tasks using face image sequences is compared in the next section.

A. Grassmann Distance Mutual Subspace Method (GD-MSM)

The Grassmann Distance Mutual Subspace Method proposed here, extends MSM in a straightforward manner, by using the Grassmann distances described in the previous section, taking into consideration all principal angles between the subspaces, instead of just using the smallest principal angle. As a result, we obtain 8 different variations, depending on which of the distances in (4) - (11) is used. Both the training and test image sequences are represented as subspaces, and the class of an arbitrary test sequence is determined as the class of the nearest training subspace, using the corresponding Grassmann distance.

B. Grassmann Discriminant Analysis (GDA)

GDA was proposed in [13] and it uses the Grassmann kernels k_P and k_{BC} in (12) - (13) in a discriminant learning framework, i.e. essentially it is a kernel discriminant analysis using Grassmann kernels. GDA applies the kernel trick to the Rayleigh quotient $L(\boldsymbol{\omega}) = \boldsymbol{\omega}^T\boldsymbol{S}_b\boldsymbol{\omega}/\boldsymbol{\omega}^T\boldsymbol{S}_\omega\boldsymbol{\omega}$, used in Linear Discriminant Analysis to find the discriminant direction $\boldsymbol{\omega}$, where $\boldsymbol{S}_b$ and $\boldsymbol{S}_\omega$ are the between-class and within-class covariances matrices. If ϕ is the feature map and $\boldsymbol{\Phi} = [\phi_1 \cdots \phi_N]$ the feature matrix of the training data (each training data is a subspace in this case), then by representing $\boldsymbol{\omega}$ as a linear combination of the feature vectors $\boldsymbol{\omega} = \boldsymbol{\Phi}\boldsymbol{\alpha}$, the Rayleigh quotient can be expressed in terms of $\boldsymbol{\alpha}$ as

$$L(\alpha) = \frac{\boldsymbol{\alpha}^T\boldsymbol{\Phi}^T\boldsymbol{S}_B\boldsymbol{\Phi}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^T\boldsymbol{\Phi}^T\boldsymbol{S}_W\boldsymbol{\Phi}\boldsymbol{\alpha}} = \frac{\boldsymbol{\alpha}^T\boldsymbol{K}(\boldsymbol{V} - \mathbf{1}\mathbf{1}^T/N)\boldsymbol{K}\boldsymbol{\alpha}}{\boldsymbol{\alpha}^T(\boldsymbol{K}(\boldsymbol{I} - \boldsymbol{V})\boldsymbol{K} + \sigma^2\boldsymbol{I})\boldsymbol{\alpha}} \quad (14)$$

where $\boldsymbol{K}$ is the kernel matrix obtained by applying one of the Grassmann kernels on the training data, $\mathbf{1}$ is an N-vector of all-ones, $\boldsymbol{V}$ is a block-diagonal matrix whose c-th block (corresponding to the c-th class) is an $N_c \times N_c$ all-ones matrix divided by N_c (the number of training samples from the c-th class), and $\sigma^2\boldsymbol{I}$ is a regularizer. GDA proceeds by finding through eigen-decomposition the values of $\boldsymbol{\alpha}$ which maximize (14), and then classification is done by nearest neighbor classification using the Euclidean distance between $F_{train} = \boldsymbol{\alpha}^T\boldsymbol{K}$ and $F_{test} = \boldsymbol{\alpha}^T\boldsymbol{K}_{test}$, where $\boldsymbol{K}_{test}$ is the kernel matrix obtained from both training and test samples.

C. Grassmann Kernel Support Vector Machine (GK-SVM)

In a similar way as the Grassmann kernels are used in conjunction with Kernel Discriminant Analysis, they can also be used in conjunction with Support Vector Machines (SVM), in a convex optimization framework. We first consider the two-class classification problem. If the training set $S = \{(\boldsymbol{Y}_1, y_1), \cdots, (\boldsymbol{Y}_N, y_N)\}$ is given, where $\boldsymbol{Y}_i$ is the matrix representations of $span(Y_i)$, corresponding to the *i*-th training image sequence, and $y_i = \{-1, 1\}$ are class labels, SVM solves the following primal optimization problem

$$\min_{\boldsymbol{\omega},\boldsymbol{\xi},b} \frac{1}{2}\boldsymbol{\omega}^T\boldsymbol{\omega} + C\sum_{i=1}^N \xi_i \quad subject\ to\ y_i(\boldsymbol{\omega}^T\phi_i + b) \geq 1 - \xi_i,\ \xi_i \geq 0 \quad (15)$$

whose dual representation (allowing the use of kernels) is given by

$$\min_{\boldsymbol{\alpha}} \frac{1}{2}\boldsymbol{\alpha}^T\boldsymbol{Q}\boldsymbol{\alpha} - \mathbf{1}^T\boldsymbol{\alpha} \quad subject\ to\ \boldsymbol{y}^T\boldsymbol{\alpha} = 0,\ 0 \leq \alpha_i \leq C \quad (16)$$

In (15) and (16), $\boldsymbol{\xi}$ is the margin slack vector, ϕ again is the feature-space transformation, C controls the trade-off between the slack variable penalty and the margin, and $\boldsymbol{Q}_{ij} = y_iy_j\boldsymbol{K}_{ij}$. The decision function is

$$sgn(\sum_{i=1}^N y_i\alpha_i\boldsymbol{K}(\boldsymbol{Y}_i, \boldsymbol{Y}_T) + b) \quad (17)$$

where $\boldsymbol{Y}_T$ is the representation of $span(Y_T)$, corresponding to a test image sequence. For multi-class classification we use the "one-against-one" approach, where for c classes, $c(c-1)/2$ binary classifiers are constructed, and classification is determined by majority voting.

D. Mutual Subspace Method (MSM)

MSM [7] corresponds to the Max Correlation distance (6) above.

E. Kernel Mutual Subspace Method (KMSM)

KMSM [8] applies kernel PCA to MSM to account for nonlinearity in the input data.

F. CLAFIC

We have implemented also CLAFIC [21], one of the earliest subspace-methods, which also represents each class c as $span(Y_c)$ through $\boldsymbol{Y}_c = [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_p]$, but rather than using distances between subspaces as a similarity measure, the angle between a test vector $\boldsymbol{x}_T$ (a single test image, or a single test pattern) and a class subspace is used

$$\cos^2 \theta = \frac{1}{||\boldsymbol{x}_T||_2^2} \sum_{i=1}^{p} (\boldsymbol{u}_i^T \boldsymbol{x}_T)^2 \tag{18}$$

Implementing CLAFIC would show whether something can be gained in terms of recognition precision, when whole image sequences or sets of image sequences are used, compared to just classifying each image individually.

Additionally, as baseline methods, we also have implemented the Eigenface method [22] based on PCA and the Fisherface method [23] based on Linear Discriminant Analysis. As CLAFIC, rather than using all the information available in a sequence or a group of sequences, these methods classify each face image separately.

IV. Experimental Results

In this section we perform experimental evaluation of the performance of the proposed methods in the context of a face recognition task. For our experiments we used the publicly available MOBIO database [18]. This data set contains a large number of face video sequences acquired primarily on a mobile phone (NOKIA N93i). Data for 152 subjects (100 males and 52 females) is available, which has been collected between August 2008 and July 2010 in six different sites from five different countries. The videos were recorded in 6 different sessions under different environmental conditions. In each session, the participants were recorded while being asked to answer a set of 21 different questions, i.e. 21 videos per session per subject are available. The questions were of different type, including both free speech and set speech. As a result of this experimental setting the video sequences contain natural facial expressions, as those accompanying natural human communication.

We have conducted two experiments, the details of which are described in the subsections below. In the first experiment, the faces were extracted from the original video sequences using the Viola-Jones face detector available from the OpenCV library. The resulting face images are not well-aligned and often contain part of the background or even some part of the face is being cropped. This experimental setting simulates the case when the data is contaminated by noise and therefore evaluates how robust the different algorithms are to noise. In the second experiment, instead of using the Viola-Jones face detector, an eye detector has been used (also available in OpenCV). Now the eyes of each subject are first detected in the original images, and then the face is cropped using the distance between the eyes. In this case all resulting face images were well aligned, and naturally this led to significant improvement in performance for all methods (this case simulates the situation when the data is clean and relatively noise-free.) One problem with this latter method for face detection is that eye detection fails more often than the Viola-Jones face detector, and therefore we generally obtain fewer faces per sequence (images in which the face detection process failed were discarded.) In all experiments the final face images were resized to 30×30 pixels.

Because of the huge size of the MOBIO database we used a subset containing the data for 100 subjects. For these subjects 10 video sequences from all 6 sessions were used, but the images in each face sequence were reduced to 25 (or even less for cases when face detection failure resulted in less than 25 faces being detected). The resulting face-only images were normalized to have zero mean and unit variance. The recognition rates reported below are the result of using a 6-fold cross validation, where in turn data from 5 sessions was used for training and the remaining one session was used as a test set.

The following eight methods were compared:
1) PCA with 1-NN (nearest neighbor) classifier; 2) LDA with 1-NN classifier; 3) CLAFIC; 4) MSM; 5) KMSM; 6) GD-MSM; 7) GK-SVM; 8) GDA. For each method, the dimension of the subspace (or the dimension of the feature space for PCA and LDA) was set experimentally to obtain the best performance. Alternatively, this could have been determined automatically by using the contribution ratio of the eigenvalues.

Additionally, for all MSM-related methods (MSM, KMSM and GD-MSM), we prepared two variants, which we call "subspace-per-category" and "subspace-per-sequence", respectively. In the subspace-per-category variant, a single subspace (or "dictionary") is learned from all training sequences for each subject, and the subspace corresponding to each test video sequence is compared to this single dictionary. In contrast, in the subspace-per-sequence variant, a multitude of subspaces (or "sub-dictionaries") are learned – a separate subspace for each training sequence, and the subspace corresponding to each test video sequence is compared to all sub-dictionaries.

A. Experiment 1 - Noisy Data

In this experiment, the faces used for training/test were detected using the Viola-Jones face detector and therefore generally contained a lot of noise and were not precisely aligned to each other.

The average recognition rates are summarized in Fig. 1, where the error bars show the standard deviation (obtained from the 6-fold cross validation) for each method. Note that for the MSM-related methods, the left bar corresponds to the subspace-per-category case, and the right bar to the subspace-per-sequence case. Also, the results for GD-MSM show the recognition rates corresponding to the best-performing among all Grassmann distances, and the results for GK-SVM and GDA show the recognition rates of the best-performing Grassmann kernels. Fig. 1 indicates that in the "noisy data case"

the Grassmann distance/kernel related methods significantly outperform the other methods, and the proposed GD-MSM and GK-SVM outperform GDA.

Fig. 2 shows the recognition rates obtained for each of the different Grassmann distances, again for both the subspace-per-category and subspace-per-sequence cases. This figure indicates that five of the Grassmann distances (the mean, projection, Binet-Cauchy, the chordal F-norm and the geodesic distances) achieve similar performance (the mean distance being slightly better than the other ones), outperforming the other three distances. It is interesting to note that the Max Correlation distance, which is used in MSM, actually performs worst. For all distances, the subspace-per-sequence variant of GD-MSM slightly outperformed the subspace-per-category one, which seems to indicate that in the case of noisy data building a single subspace from all training sequences might be disadvantageous.

Fig.3 shows the recognition rates for the kernel-based methods, GK-SVM and GDA, giving the recognition rates for each of the Grassmann kernels.The proposed GK-SVM outperforms GDA in this experiment, and the results are stable for both kernels. GDA performs well with the Binet-Cauchy kernel, but recognition rate drops sharply when the projection kernel is used.

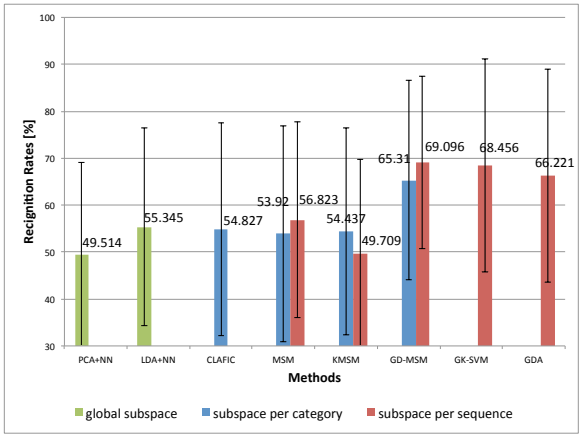

Fig. 1. Recognition rates for all methods in the noisy data case (experiment 1, face detector used)

B. Experiment 2 - Clean Data

In this experiment, the faces used for training/test were detected from the original face video sequences using the OpenCV eye detector, and then automatically put into alignment by utilizing the position of the detected eyes. In this way we were able to obtain a relatively clean data set, which allows to compare all methods under the best possible conditions.

Fig. 4 shows the average recognition rates for all methods. Here again best performance is achieved by GD-MSM followed by GK-SVM and GDA, i.e. again the Grassmann distance/kernel related methods outperform the other methods, although the difference is not as impressive as in the noisy data

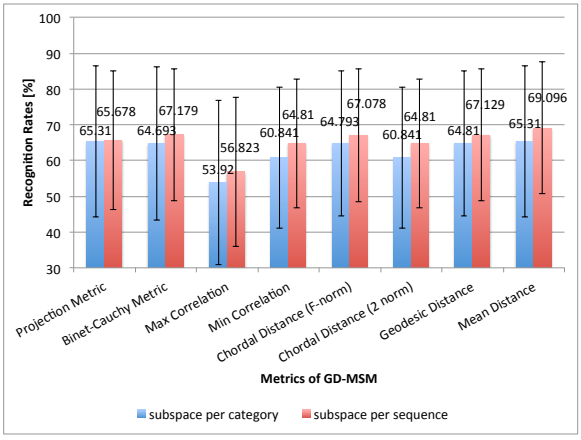

Fig. 2. Recognition rates for the different Grassmann distances (experiment 1)

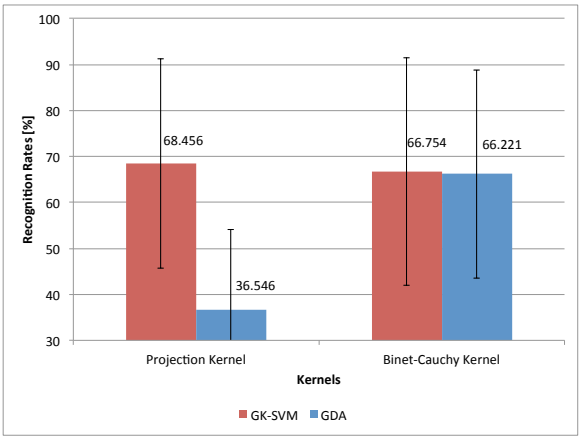

Fig. 3. Recognition rates for the different Grassmann kernels used in GK-SVM and GDA (experiment 1)

case in Fig. 1. It is interesting to note that here the subspace-per-category variant achieves better performance than the subspace-per-sequence one, which indicates that when the faces are well-aligned and relatively clean from noise it is advantageous to merge all training data for each subject into a single subspace.

Fig. 5 shows the recognition rates for the different Grassmann distances. Here also the superiority of the subspace-per-category variant over the subspace-per-sequence one is obvious. Now best performance is obtained for the geodesic distance, while the Max Correlation distance again performs worst. The recognition rates for the different Grassmann kernels are given in Fig. 6, and again best performance is achieved for GK-SVM with the projection kernel. For the clean data set, GDA's performance with the projection kernel improves, however still it is way below GK-SVM, or GDA with the Binet-Cauchy kernel.

Overall, the results from both experiments show that the Grassmann distance/kernel related methods significantly outperform both the subspace-related methods which do not use the Grassmann metric (CLAFIC, MSM and KMSM) and the

global-subspace methods like Eigenface (PCA) and Fisherface (LDA). Additionally, when Grassmann distances are used, the experiments show that using a subspace-per-sequence representation is advantageous in the case when the data contains a lot of noise, while for relatively clean datasets the subspace-per-category representation is more appropriate.

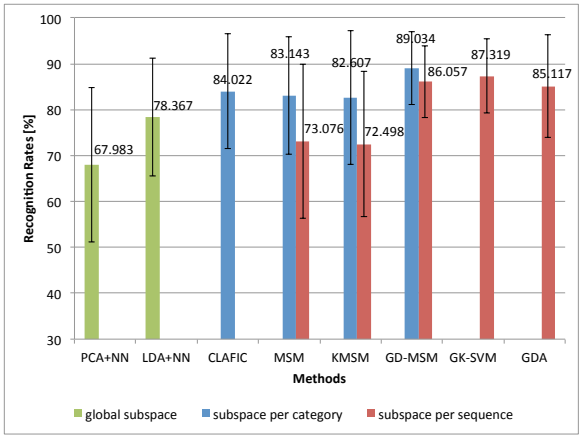

Fig. 4. Recognition rates for all methods in the clean data case (experiment 2, eye detector used)

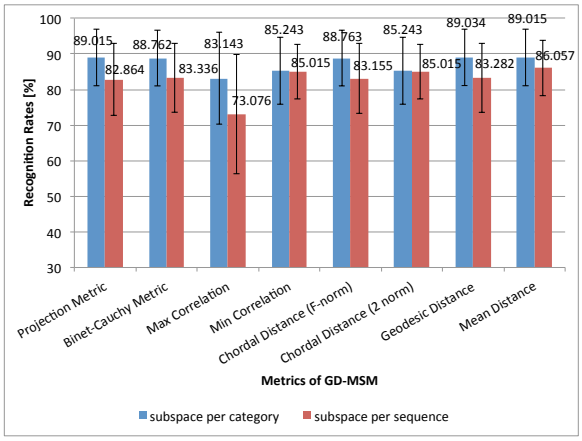

Fig. 5. Recognition rates for the different Grassmann distances (experiment 2)

V. Conclusion

In this paper we have proposed two novel methods. First, we extend the Mutual Subspace Method [7], so that rather than considering only the cosine of the smallest principal angle (which geometrically is not a good measure for subspace similarity and also not optimal for recognition), Grassmann distances are used to define the similarity between image sequences. Second, we apply the projection and Binet-Cauchy Grassmann kernels in a convex optimization scheme, using the Support Vector Machine (SVM). The resulting methods, GD-MSM and GK-SVM were experimentally compared with several related methods on a large database of videos containing face sequences from 100 subjects with expression changes related to a natural conversation setting.

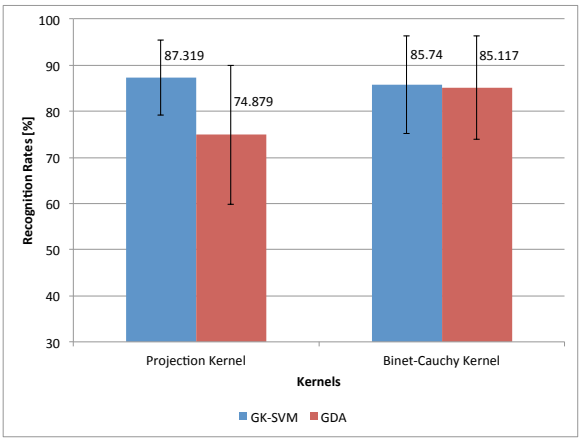

Fig. 6. Recognition rates for the different Grassmann kernels used in GK-SVM and GDA (experiment 2)

The experimental results show that for both noisy and clean datasets, the proposed methods significantly outperform subspace-related methods which do not use the Grassmann metric (like CLAFIC, MSM and KMSM) and global-subspace learning methods like Eigenface and Fisherface.

Additionally, we studied the effect of combining all available image sequences into one large dictionary (learning a common subspace from all available image sequences for each category), in comparison with using multiple smaller subspaces (i.e. representing each category by several different subspaces, where each subspace is formed from image sequences taken under different conditions). The experiments showed that using a subspace-per-sequence representation is advantageous in the case when the data contains a lot of noise, while for relatively clean datasets the subspace-per-category representation is more appropriate.

As a further work, it would be interesting to apply the Grassmann metric-based methods proposed here to other problems where the data can be represented as sets of vectors, and for which the subspace representation is more natural and efficient than a vector representation.

References

[1] G. Shakhnarovich, J.W. Fisher, and T. Darrel, "Face Recognition from Long-Term Observations," *in Proc. European Conf. Computer Vision (ECCV),* pp. 851-868, 2002.

[2] O. Arandjelovic, G. Shakhnarovich, J. Fisher, R. Cipolla, "Face Recognition with Image Sets Using Manifold Density Divergence," *in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR),* pp. 581-588, 2005.

[3] S. Satoh, "Comparative Evaluation of Face Sequence Matching for Content-Based Video Access," *in Proc. Int. Conf. Automatic Face and Gesture Recognition,* pp. 163-168, 2000.

[4] B. Raytchev and H. Murase, "Unsupervised Recognition of Multi-View Face Sequences Based on Clustering with Attraction and Repulsion," *Computer Vision and Image Understanding,* Vol. 91, No. 1-2, pp. 22-52, 2003.

[5] A. Bjorck and G.H. Golub, "Numerical Methods for Computing Angles between Linear Subspaces," *Math Computation,* vol.27, no. 123, pp. 579-594, 1973.

[6] G.H. Golub and C.F. van Loan, *Matrix Computations,* Johns Hopkins University Press, 3rd edition.

[7] Yamaguchi, K. Fukui, and K. Maeda, "Face Recognition Using Temporal Image Sequence," *in Proc. Int. Conf. Automatic Face and Gesture Recognition,* pp. 318-323, 1998.

[8] H. Sakano and N. Mukawa, "Kernel mutual subspace method for robust facial image recognition," *in Proc. Int. Conf. on Knowledge-Based Intell. Eng. Sys. And App. Tech.,* pp. 245-248, 2000.

[9] K. Fukui and O. Yamaguchi, "Face Recognition Using Multi-Viewpoint Pattern for Robot Vision," *in Proc. Int. Symp. Robotics Research,* pp. 192-201, 2003.

[10] M. Nishiyama, O. Yamaguchi, and K. Fukui, "Face Recognition with the Multiple Constrained Mutual Subspace Method," *in Proc. 5th International Conference on Audio- and Video-based Biometric Person Authentication (AVBPA),* pp. 71-80, 2005.

[11] T.-K. Kim, O. Arandjelovic, and R. Cipolla, "Learning over Sets Using Boosted Manifold Principal Angles (BoMPA)," *in Proc. British Machine Vision Conf.,* pp. 779-788, 2005.

[12] T.-K. Kim, J. Kittler, and R. Cipolla, "Discriminative learning and recognition of images set classes using canonical correlations," *in IEEE Trans. Pattern Anal. Mach. Intell. (PAMI),* 29, 1005-1018, 2007.

[13] J. Hamm and D. D. Lee, "Grassmann discriminant analysis: A unifying view on subspace-based learning," *in Proc. 25th Int. Conf. on Machine Learning,* pp. 376-383, 2008.

[14] Y. C. Wong, "Differential geometry of Grassmann manifolds," *in Proc. of the Nat. Acad. of Sci.,* Vol. 57, pp. 589-594, 1967.

[15] A. Edelman, T.A. Aris and S.T. Smith, "The geometry of algorithms with orthgonality constrains," *SIAM J. Matrix Anal. Appl.,* 20 (2), pp. 303-353, 1998.

[16] T. Wang and P. Shi, "Kernel Grassmannian distances and discriminant analysis for face recognition from image sets," *Pattern Recognition Letters,* Vol. 30, pp. 161-165, 2009.

[17] C.M.Bishop, *Pattern recognition and Machine Learning,* Springer-Verlag, 2006.

[18] C. McCool and S. Marcel, "MOBIO Database for the ICPR 2010 Face and Speech Competition," *an IDIAP Research Institute Communication,* Idiap-Com-02-2009, 2009.

[19] J.S. Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis,* Cambridge University Press, 2004.

[20] C. Cortes and V. Vapnik, "Support Vector Networks," *Machine Learning,* 20, pp. 273-297, 1995.

[21] S. Watanabe and N. Pakvasa, "Subspace Method of Pattern Recognition," *1st, Int. Joint Conf. on Pattern Recognition,* pp. 25-32, 1973.

[22] M. Turk and A. Pentland, "Eigenfaces for Recognition," *Journal of Congnitive Neuroscience,* vol. 3, no. 1, pp. 71-86, 1991.

[23] P.N.Belhumeur, J.P.Hespanha, and D.J.Kriegman, "Eigenfaces versus Fisherfaces: Recognition Using Class Specific Linear Projection," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 711-720, 1997.