CMML: a New Metric Learning Approach for Cross Modal Matching

Alexis Mignon, Frédéric Jurie

CMML: a New Metric Learning Approach for Cross Modal Matching

Alexis Mignon and Frédéric Jurie

GREYC, CNRS UMR 6072, Université de Caen Basse-Normandie, France

first_name.last_name@unicaen.fr

Abstract. This paper proposes a new approach for Cross Modal Matching, *i.e.* the matching of patterns represented in different modalities, when pairs of same/different data are available for training (e.g. faces of same/different persons). In this situation, standard approaches such as Partial Least Squares (PLS) or Canonical Correlation Analysis (CCA), map the data into a common latent space that maximizes the covariance, using the information brought by positive pairs only. Our contribution is a new metric learning algorithm, which alleviates this limitation by considering both positive and negative constraints and use them efficiently to learn a discriminative latent space. The contribution is validated on several datasets for which the proposed approach consistently outperforms PLS/CCA as well as more recent discriminative approaches.

1 Introduction

This paper addresses Cross Modal Matching (CMM), which is the task of predicting whether a pair of data points, coming from two different modalities, represents the same object. Many computer vision tasks are related to CMM, such as face recognition (e.g. photos vs. sketches, high-resolution photos vs. low-resolution photos, front vs. profile, etc.) or person re-identification across multiple cameras. In addition to computer vision, many other fields are interested in CMM, since when observations comes from different devices or are represented in different ways (e.g. textual description vs. image) a link has to be established between them if they are to be compared.

As explained in the next section, one can distinguish two main ways for addressing CMM. First, mapping (or synthesizing) the data represented in one modality onto the other modality makes all the data comparable. Second, it is also possible to learn a common latent space, different from the two initial representation spaces, and independently project the data, from the original (distinct) spaces into the new space, making them directly comparable. The work presented here follows the second way.

More specifically, we are interested in the tasks for which we have training data that can be used to optimally relate the two modalities. We consider the case where training data are given as two sets of data points, in two different modalities, related through a set of pairwise matching/non-matching constraints.

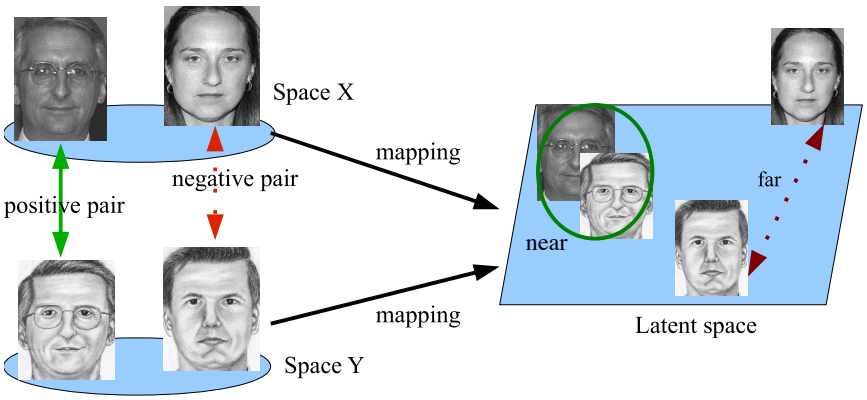

Fig. 1. Cross Modal Matching: Two sets of data points in two different spaces X (photos) and Y (sketches) are related through a set of pairwise positive/negative constraints. We learn a common latent space in which distances reflect the constraints.

E.g. for face verification with photos and sketches, training pairs are sets photo-sketch pairs of same person, as well as photo-sketch pairs of faces of different persons, as illustrated Fig.1. We refer to them later as positive and negative constraints respectively. Note that the pairwise constraint setting is very general since it includes, for instance, problems in which class labels are known. In this case pairwise constraints are known for all possible pairs.

Finding common latent spaces for CMM is an important issue and has therefore received a lot of attention. Several very popular approaches have been proposed, such as Canonical Correlation Analysis (CCA) [1] and Partial Least Squares (PLS) [2]. All of them aim at finding two projection matrices that lead to maximal covariance in the latent space, under the positive constraints given by the training data. For instance, CCA computes two projections that minimize the distance between positive pairs in the latent space.

One severe drawback of these approaches is that they ignore the negative constraints, *i.e.* the constraints defined by pairs of different objects. However, for many tasks, those negative constraints are available and using them can definitely bring some benefits in terms of performance. Some recent approaches have used both constraints [3, 4] but they suffer from different limitations (see next section). Overcoming these limitations is precisely the aim of this paper.

This paper proposes a metric learning approach that takes advantage of both positive and negative constraints, for cross modal matching. The key idea is to learn a latent space shared by both modalities in which similar objects are closer than different objects. The proposed approach is validated on 3 different datasets for different cross-modal face recognition tasks: on CUFSF [5] for face vs. sketch recognition, on Multi PIE [6] for cross pose face recognition, and on LFW [7] for face verification using different face descriptors. These experiments demonstrate that using negative constraints improves the performance of cross-

modal matching, and show that the proposed algorithm consistently outperforms existing competing latent space learning algorithms such as CCA and PLS.

2 Related works

Considered as being an important problem, Cross Modal Matching (CMM) has already received a lot of attention in the literature. Works can be organized into two domains: on one side those that map data of the first modality into the space of the second, and, on the other side, those building a common latent space in which both modalities are projected before they can be compared.

When data represented in the first modality are mapped into the second modality, the synthesized data can then be directly compared into the second space. For instance, [8] maps near infrared images to visible images; in addition of allowing the direct comparison of the two modalities, a human operator can also analyze more easily the images. In the same way, [9] and [5] synthesize face sketches from photographs to perform direct comparison between sketches.

When the two modalities are close enough, applying specific filters can drastically reduce the differences between them and eventually make them directly comparable [10, 11]. However, this strategy can be used only in very specific contexts.

Despite the good results obtained by synthesis-based approaches, they are usually 'task specific' in the sense that the synthesis framework needs to be redefined for each new problem, if possible at all. Consequently, these methods do not generalize well to new problems. It explains why most of the works, including ours, focus on defining common latent projection spaces.

The latent representation approach assumes that observations in both modalities can be explained by a reduced set of common latent variables. They often rely on popular statistical tools such as Canonical Correlation Analysis (CCA) [1] or Partial Least Squares (PLS) [2], which aim at finding the latent representation by maximizing covariance in the latent space. The main difference between these two methods lies in the constraints controlling the quality of the reconstruction in the original space [12]. PLS was recently used in [13] for different cross-modal matching tasks (face sketch/photo recognition, low/high-resolution matching, cross-pose recognition) for which it was shown to be competitive with other ad-hoc methods [9, 11]. CCA was used in [14] to recognize occluded faces by matching non-corresponding parts of the faces (*e.g.* eyes and nose). In [15], sketches and face images are encoded so that the two representations of the same training person are identical. This is achieved by learning a tree structure in which nodes project a subset of the sketch and photograph features into a common space using CCA. The method was reported to give state-of the art results on the CHUK Face Sketch FERET database (CUFSF).

These covariance-based methods, by construction, ignore the negative constraints, *i.e.* the constraints given by pairs of non-matching points. Addressing this issue, [16] suggests to apply linear discriminant analysis (LDA) in each input space, and learn the common latent space with CCA, in the context of

infra-red/visible face matching. [17] adopts a similar strategy but introduces coupling terms to perform the two LDA simultaneously. These two methods use the discriminative information in the initial input spaces, while it would be more appropriate to integrate it directly with the construction of the latent space. This is precisely what [18] does, by deriving an equivalent of LDA directly in the latent space, in such a way that projection and discriminant analysis are performed simultaneously. The common point between this method and the two previous ones is that they all rely on LDA and thus require the data to be fully annotated i.e. classes and class memberships must be explicitly defined, to be able to compute the within class scatter matrices. This is often not possible as in general, only pairwise matching/non-matching information is available, without any class labeling.

Alternatively, Common Discriminant Feature Extraction (CDFE) [3], as well as the very similar Discriminant Mutual Subspace Learning (DMSL) [4], can deal with arbitrary sets of pair-wise constraints directly in the latent space. They both aim at minimizing distances between positive pairs while maximizing distances between negative pairs. CDFE also includes some regularization terms to enforce distance consistency between the input spaces and the latent space. One severe limitation of this line of works lies in the objective function used to learn the latent space. While intuitive, maximizing distances given by negative pairs and minimizing distances given by positive pairs may not be relevant, if two points of a negative pair are already farther than any pair of points of positive constraints, pushing them even farther – which is definitively going to happen since distances corresponding to positive constraints are bounded by zero – does not improve discrimination and can even lead to over-fitting. We believe the important information lies in the 'margin', *i.e.* the range of distances for which negative and positive constraints are close, and not in the farthest points.

Our approach bears similarity with [3] and [4] since we use both positive and negative constraints to learn a discriminative latent space. However, our Cross Modal Metric Learning (CMML) method addresses the problem from the opposite point of view, in the sense that it penalizes inappropriate distances instead of favoring expected ones. More precisely, we introduce a new objective function such that distances corresponding to positive pairs are penalized when they are greater than a threshold while those corresponding to negative pairs are penalized when they are smaller than the same threshold. This allows to treat symmetrically positive and negative constraints, since the contribution of pairs with appropriate distances become quickly insignificant. From this point of view, CMML adopts an objective function which is in the spirit of the popular Support Vector Machines or Logistic Regression.

Learning a subspace in which distances reflect a set of similarity constraints is not new in the monomodal case. Indeed methods such as Neighborhood Component Analysis (NCA) [19], Large Margin Component Analysis (LMCA) [20] and Pairwise Constrained Component Analysis (PCCA) [21] precisely address this problem. Among these methods PCCA is the only one able to deal with pairwise constraints and CMML can be seen as the generalization of PCCA to

cross-modal settings. To our knowledge, this way of extending traditional metric learning methods to cross-modal problems has not been investigated before.

3 Cross modal metric learning (CMML)

As explained before, our goal is to learn a parametric distance function according to a set of pairwise matching/non-matching constraints, allowing the comparison of data lying in two different spaces (modalities). Given two sets of data points $S_x = \{\boldsymbol{x}_i\}_{i=1}^{N_x}$ and $S_y = \{\boldsymbol{y}_j\}_{j=1}^{N_y}$ coming from the two different spaces $\mathcal{X}$ and $\mathcal{Y}$, we want to learn two transformations f_A and g_B, parametrized by A and B, which independently map the points from $\mathcal{X}$ and $\mathcal{Y}$, to a k-dimensional Euclidean space $\mathcal{Z}$ in which the distance is measured as:

$$D^2_{A,B}(\boldsymbol{x}, \boldsymbol{y}) = \|f_A(\boldsymbol{x}) - g_B(\boldsymbol{y})\|^2 \tag{1}$$

Let $\mathcal{C}$ be a set of N_c constraints given as triplets $(i, j; l)$ where i (resp. j) is the index of the i-th (resp. j-th) data sample in S_x (resp. S_y). l is a label indicating if the elements of the pair (i, j) match ($l = 1$) or do not match ($l = -1$). Two elements match if the two modalities represent the same object (e.g. the same face). We will refer to these pairs respectively as positive and negative pairs.

To learn the best parameters A and B, we build a loss function penalizing distances of positive pairs greater than a threshold (arbitrarily set to 1), and distances of negative pairs smaller than the same threshold. For a $(i, j; l)$ constraint, the loss function is:

$$\ell_\beta\left(l(D^2_{A,B}(\boldsymbol{x}_i, \boldsymbol{y}_j) - 1)\right) \tag{2}$$

We use the generalized logistic loss function $\ell_\beta(x) = \log(1 + e^{\beta x})/\beta$ instead of the more straightforward hinge loss function $h(x) = \max(0, x)$ to avoid numerical instabilities due to the discontinuity of the gradient during the optimization process. $\ell_\beta(x)$ can be seen as a smooth approximation of the hinge loss function with the parameter β controlling the *smoothness* of the approximation ($\lim_{\beta\to\infty} \ell_\beta(x) = h(x)$).

Summing the loss of Eq. (2) for all constraints in $\mathcal{C}$, optimal values of A and B can be obtained by minimizing the following objective function :

$$A^*, B^* = \arg\min_{A,B} \mathcal{L}(A, B) \tag{3}$$

$$\mathcal{L}(A, B) = \sum_{(i,j,l)\in\mathcal{C}} \ell_\beta(l(D^2_{A,B}(\boldsymbol{x}_i, \boldsymbol{y}_j) - 1)) \tag{4}$$

In the following, we address the common case where $\mathcal{X} = \mathbb{R}^{d_x}$ and $\mathcal{Y} = \mathbb{R}^{d_y}$ are two Euclidean spaces, and $f_A(\boldsymbol{x}) = A\boldsymbol{x}$ and $g_B(\boldsymbol{y}) = B\boldsymbol{y}$ are two linear transformations parametrized by the $k \times d_x$ (respectively $k \times d_y$) dimensional matrix A (respectively B). An extension to the non-linear case for arbitrary Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$ is then proposed, through the use of the *kernel trick*.

3.1 Optimization

The optimization is performed using a simple gradient descent scheme, the gradient being:

$$\frac{\partial \ell_\beta(A,B)}{\partial A} = 2 \sum_{i,j,l \in \mathcal{C}} l\sigma_\beta(l(\|A\boldsymbol{x}_i - B\boldsymbol{y}_j\|^2 - 1)(A\boldsymbol{x}_i - B\boldsymbol{y}_j)\boldsymbol{x}_i^T \tag{5}$$

$$\frac{\partial \ell_\beta(A,B)}{\partial B} = -2 \sum_{i,j,l \in \mathcal{C}} l\sigma_\beta(l(\|A\boldsymbol{x}_i - B\boldsymbol{y}_j\|^2 - 1)(A\boldsymbol{x}_i - B\boldsymbol{y}_j)\boldsymbol{y}_j^T \tag{6}$$

The loss function $\mathcal{L}(A,B)$ is clearly not convex but our optimization problem can be reformulated as a low rank factorization of a convex function. By denoting $\boldsymbol{z}_c^T = (\boldsymbol{x}_i^T, \boldsymbol{y}_j^T)$ the the concatenation of $\boldsymbol{x}_i$ and $\boldsymbol{y}_j$ corresponding to the the c-th constraint of $\mathcal{C}$, and $L = [A\ (-B)]$ the $k \times (d_x + d_y)$ dimensional matrix where each row is the concatenation of the corresponding rows of A and B, and by denoting $M = L^T L$, $\mathcal{L}(A,B)$ as:

$$\mathcal{L}(A,B) = \sum_{c=1}^{N_c} \ell_\beta(l_c(\boldsymbol{z}_c^T L^T L \boldsymbol{z}_c - 1)) \tag{7}$$

$$= \sum_{c=1}^{N_c} \ell_\beta(l_c(\boldsymbol{z}_c^T M \boldsymbol{z}_c - 1)) \tag{8}$$

Due to the convexity of $\ell_\beta(x)$, Eq. (8) is a convex function of M. From [22] and [23], we can assert that minimizing Eq. (7) with respect to L, through a gradient based optimization scheme is equivalent to minimizing Eq. (8) with respect to M if $k \geq \text{rank}(M^*)$, M^* being the actual optimum of Eq. (8).

3.2 Kernel CMML

Comparing descriptions from different spaces through linear combinations may have no sense in general. Fortunately, the algorithm can be easily generalized to any pair of Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$ by rewriting forumulæ so that only the kernel matrices appear.

Let X (respectively Y) be the matrix where each column is a vector from S_x (respectively S_y) and $K_x = X^T X$ (resp. $K_y = Y^T Y$) the corresponding kernel matrix. By parameterizing A and B as $A = \hat{A}X^T$ and $B = \hat{B}Y^T$, $A\boldsymbol{x}_i$ (resp. $B\boldsymbol{y}_j$) becomes $\hat{A}X^T\boldsymbol{x}_i = \hat{A}\boldsymbol{k}_{xi}$ (resp. $\hat{B}\boldsymbol{k}_{yj}$), where $\boldsymbol{k}_{xi}$ (resp. $\boldsymbol{k}_{yj}$) is the i-th row (resp. j-th) of K_x (resp. K_y). So Eq. (4) becomes:

$$\mathcal{L}(A,B) = \mathcal{L}'(\hat{A},\hat{B}) = \sum_{i,j,l \in \mathcal{C}} \ell_\beta(l(\|\hat{A}\boldsymbol{k}_{xi} - \hat{B}\boldsymbol{k}_{yj}\|^2 - 1)) \tag{9}$$

The kernel formulation makes the algorithm formally independent from the representation chosen, and thus, should be considered as the standard form of the algorithm.

Using the gradient formula from eq. (5) and defining the shortcut $\sigma_t = \sigma_\beta(l(\|A_t\boldsymbol{x}_i - B_t\boldsymbol{y}_j\|^2 - 1))$, we derive the update rule for the gradient descent scheme from the linear case:

$$A_{t+1} \leftarrow A_t + 2\eta \sum_{i,j,l\in\mathcal{C}} l\sigma_t(A_t\boldsymbol{x}_i - B_t\boldsymbol{y}_j)\boldsymbol{x}_i^T \tag{10}$$

$$\hat{A}_{t+1}X^T \leftarrow \hat{A}_tX^T + 2\eta \sum_{i,j,l\in\mathcal{C}} l\sigma_t(\hat{A}_tX^T\boldsymbol{x}_i - \hat{B}_tY^T\boldsymbol{y}_j)\boldsymbol{x}_i^T \tag{11}$$

where η is the adaptation step. Multiplying both sides by XK_x^{-1}, leads to:

$$\hat{A}_{t+1} \leftarrow \hat{A}_t + 2\eta \sum_{i,j,l\in\mathcal{C}} l\sigma_t(\hat{A}_t\boldsymbol{k}_{xi} - \hat{B}_t\boldsymbol{k}_{y_j})\boldsymbol{k}_{xi}^TK_x^{-1} \tag{12}$$

Which is equivalent to use the gradient in eq. (5) right multiplied by K_x^{-1}. This acts like a preconditioning of the gradient in eq. (5) and leads to better convergence rates, as shown by [24] in the context of SVM classification.

By observing that $\boldsymbol{k}_{xi}^TK_x^{-1} = \boldsymbol{e}_i^T$, where $\boldsymbol{e}_i$ is the vector with the i-th element equal to 1 and all others equal to 0, it appears that only one column of A needs to be updated for each pair instead of the full matrix in the linear case (Eq. (10)). This leads to further gain in convergence speed.

4 Experiments

In this section we validate our contribution, the Cross Modal Metric Learning algorithm (CMML), through different multimodal experiments on three different databases: CMU Multi PIE , CHUK Face Sketch FERET (CUFSF) and Labeled Faces in the Wild (LFW). We report the performance of our approach on these datasets and provide comparisons with the related approaches that can work with pairwise constraints (*i.e.* Canonical Correlation Analysis (CCA) [1], Partial Least Square (PLS) [2] and Common Discriminant Feature Extraction (CDFE) [3, 4]).

We first describe the databases and the associated experimental protocols, then give some implementation details and, finally, present the results.

4.1 Databases and evaluation protocols

Multi PIE. CMU Multi PIE (for Pose, Illumination and Expression) [6] contains face images taken under different view points and hence having different head poses, illumination changes and varying expressions. Fig. 3 shows the locations and labels of the cameras. Four different acquisition sessions were performed over 5 months.

In our experiments, we use the first session which is the largest one, and keep only images with frontal illumination and neutral expression. There are 249 persons depicted in this session, viewed under 15 different poses each. Images were aligned using eyes, nose and mouth corners. A tight crop was also applied to remove most of the background. Then, uniform Local Ternary Pattern descriptors (LTP) [25] were extracted over a uniform grid. Since the size of the so cropped

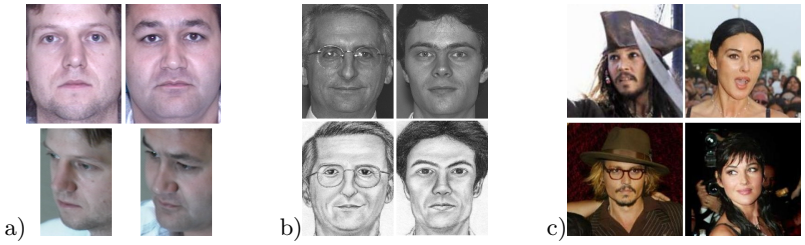

Fig. 2. Sample images from Multi PIE, CUFSF and LFW datasets. a) Multi PIE: cropped images from cameras "05_1" (top row) and "08_1" (bottom row) (see Fig. 3 for more precision on camera positions). b) CUFSF: two cropped faces (top row) and their corresponding sketches (bottom row). c) LFW: two positive pairs.

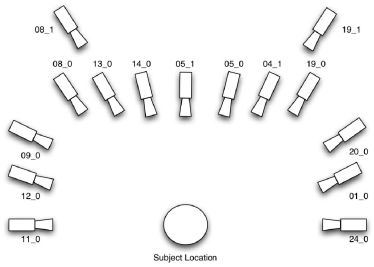

Fig. 3. Multi PIE: camera labels/locations for capturing images. 13 cameras were located at head height, spaced every 15°. Two additional cameras ("08_1" and "19_1") were above the subject, simulating video surveillance cameras. Figure reproduced from [6].

images is different for each pose (see Fig. 4.1-a for an illustration), descriptors dimensionality are also pose dependent (from 5664 to 9440 components).

The train set is obtained by randomly choosing 149 persons among the 249, the 100 remaining ones being assigned to the test set. Negative constraints are obtained by producing pairs containing two different persons, while positive constraints are pairs of the same person. Comparisons between the different methods are performed using the same amount of positive and negative training pairs. However, as negative constraints can be generated without limit, we also studied to what extend this amount affects the results, by randomly generating negative pairs.

Algorithms are trained for each possible pairs of poses ($15 \times 14 = 210$ pairs), meaning there is one different latent space for each pair of camera positions.

Performance is evaluated by two different measures. The first one evaluates the performance in a face retrieval context. Given an image in one view (the probe) we determine the closest match in a set of images in another view (the gallery). The classification is correct if the closest gallery image depicts the same person as the probe. The average number of correct matches over the test set is reported as the *nearest neighbor accuracy*. The second measure considers a face verification context. The test set contains as many positive pairs as negative

ones. Each person is involved in one positive and one randomly chosen negative pair. Pairs are classified as positive if the corresponding distance is lower than a threshold chosen as the median distance between elements of pairs in the test set. The rate of correctly classified pairs corresponds to the Equal Error Rate (EER) of the corresponding ROC curve, and is therefore reported as the *accuracy at equal error rate*.

The overall process is repeated 10 times with different sets of random splits and negative pairs and the mean and standard deviation are reported.

CUFSF. CHUK Face Sketch FERET (CUFSF) [5] includes 1194 persons from the FERET database [26]. For each person, a sketch has been drawn by an artist, with the intention of making part of their appearance more noticeable than it really is. Because some images of the original FERET database have been removed from the now available (extended) version of FERET, only 860 photo/sketch pairs are nowadays available. As for the experiments on Multi PIE, we extracted LTP descriptors on aligned and cropped images, using the annotation provided with the database. The final descriptors are 9440 dimensional vectors for both modalities.

The algorithms are trained using a subset of 500 randomly chosen persons, the remaining 360 persons being used for testing.

The experiments were made following exactly the same protocols as for Multi PIE, and the performance given by the nearest neighbor accuracy and the accuracy at equal error rate (see previous section).

LFW. Despite Labeled Faces in the Wild (LFW) [7] is not intrinsically multimodal, it has been extensively used in the past, giving the opportunity to compare the results with existing works. LFW contains more than 13,000 images of faces collected from the web. 1680 of the pictured people are represented in two or more photographs. The database is made of 10 sets of 300 positive pairs and 300 negative pairs. People cannot be represented in more than one set. The performances are reported as the average classification accuracy of a 10-fold cross validation.

We have introduced cross-modality by representing each face of a face pair with a different type of descriptor. On one hand, we use the descriptors used in the previous experiments, *i.e.* LTP computed on the aligned version of the database [27]. On the other hand, we also represent face image with the face descriptors used by [28] (downloaded from author's web page). They are SIFT [29] features extracted on 9 face key-points (eye corners, nose corners, nose tip and mouth corners) at 3 different scales.

The dimensionality of the two descriptors is respectively 14160 and 3456. All the experiments done on LFW uses pairs of these very different descriptors, supposed to represent two different modalities that cannot be compared directly.

As we use the *restricted* setting of LFW [7], labels are given only for pairs and we don't know the identity of the persons involved. Therefore, the nearest neighbor accuracy would not make sense and we report only the accuracy at equal error rate (see Section 4.1), following LFW's protocol.

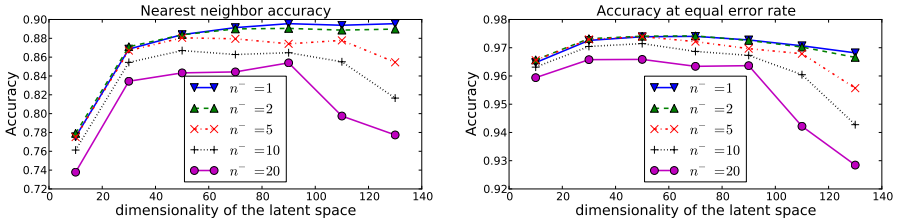

Fig. 4. CMML's average accuracies on Multi PIE, for different ratios n^- of negative/positive training pairs and for different dimensionalities of the latent space.

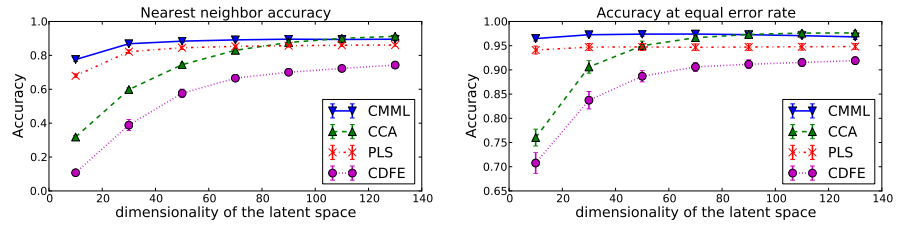

Fig. 5. Multi PIE : average accuracies of CMML, CCA, PLS and CDFE, with $n^- = 1$.

Implementation details. Since all the descriptors used in our experiments are histograms, we use the χ^2-RBF kernel to compare them:

$$D^2_{\chi^2}(\boldsymbol{x}, \boldsymbol{y}) = \sum_i \frac{(x_i - y_i)^2}{x_i + y_i} \tag{13}$$

$$K(\boldsymbol{x}, \boldsymbol{y}) = e^{-\alpha D^2_{\chi^2}(\boldsymbol{x}, \boldsymbol{y})} \tag{14}$$

We have observed that $\alpha = 2$ gives good results on average and used it for all the presented experiments. Furthermore, for making the comparisons of the different method possible, all of them have been used in their kernel version with the χ^2-RBF kernel.

For CMML the smoothness parameter β was set to 3 for all the experiments (in practice we have observed that this value does not affect the performance a lot, explaining why we have used the same value for all the experiments). For CCA and CDFE the regularization parameters are cross-validated for each experiment, ensuring to give optimal results.

4.2 Results

Results on Multi PIE. We first evaluate the performance of our method (CMML) for different number of negative training pairs and for different sizes of the latent space. As explained before, we use two different measures of performance: (a) Nearest neighbor accuracy and (b) accuracy at Equal Error Rate (EER) (see Section 4.1). Fig. 4 shows the performance of CMML, using all pairs of possible poses, for different sizes of the latent space and for different ratios of

Method \ Gallery	11_0	12_0	09_0	08_0	08_1	13_0	14_0	05_0	04_1	19_1	19_0	20_0	01_0	24_0
CMML	0.75	0.82	0.89	0.95	0.85	0.94	1.00	1.00	0.99	0.82	0.92	0.89	0.76	0.64
CCA	0.28	0.31	0.30	0.52	0.26	0.43	0.53	0.95	0.94	0.46	0.47	0.38	0.73	0.28
PLS	0.50	0.72	0.80	0.93	0.69	0.93	1.00	1.00	0.99	0.60	0.88	0.76	0.67	0.50
CDFE	0.14	0.28	0.26	0.35	0.15	0.43	0.58	0.51	0.31	0.14	0.33	0.21	0.15	0.18

Table 1. Nearest neighbor accuracy for pairs made of one "05_1" image (frontal view) and images with different poses (one different pose per column), for CMML, CCA and PLS,. The dimensionality of the latent space is 30 and $n^- = 1$. See Fig. 3 camera labels.

negative/positive training pairs ratio. $n^- = k$ means there are k times more negative pairs than positive pairs in the training set. $n^- = 1$ gives the best result. Performance decreases when this ratio increases, and the drop seems even more important for high-dimensional latent spaces. This is probably because of two factors: i) when n^- increases, too much weight is given to negative constraints with respect to positive ones and ii) over-fitting.

Note that in these experiments, negative pairs were generated randomly. We also tried to generate them using heuristics like considering for each person the closest negative match, but this approach tends decrease the performances, most probably because it enforces over-fitting.

We then compare CMML with its competitors (CCA, PLS and CDFE). Fig. 5 shows the performance of the different methods, averaged over all pairs of possible poses, as a function of the dimensionality of the latent space (for $n^- = 1$). For low dimensional latent spaces, CMML and PLS clearly outperform CCA and CDFE, with a clear advantage for CMML, which remains consistently higher than PLS. The performance of CMML and PLS are quite stable with respect to the dimensionality while CCA and CDFE performs better when the dimensionality gets higher. CDFE performs consistently worse than all other methods, while CCA outperforms CMML when the number of dimensionality is larger than about one hundred.

Tab. 1 reports the nearest neighbor accuracy for pairs made of i) one front view (given by camera "5_01") and ii) another image coming from another camera view. Each column reports the accuracy for the corresponding camera view. Therefore, the first column gives the accuracy for pairs made of "5_01" and "11_0" images. Views are sorted from left to right accordingly with their spatial organization, as illustrated by Fig. 3. Consequently, profile views (cameras "11_0" and "24_0") lie at the right and left-hand sides. Please, note that for these experiments, the size of the latent space is 30. As expected, the matching is easier for near-frontal views (cameras "14_0" and "5_0") for which CMML and PLS give perfect matching. As we can see by looking at Table 1, CMML outperforms all other methods.

Results on CUFSF. Similar experiments have been done on CUFSF. Fig. 6 shows that CMML outperforms all other methods by a large margin. As for Multi PIE the performances are quite steady with the dimensionality of the latent space. While the accuracy of the nearest neighbor increases to saturate around 70 dimensions, the accuracy at equal error rates remains roughly the same

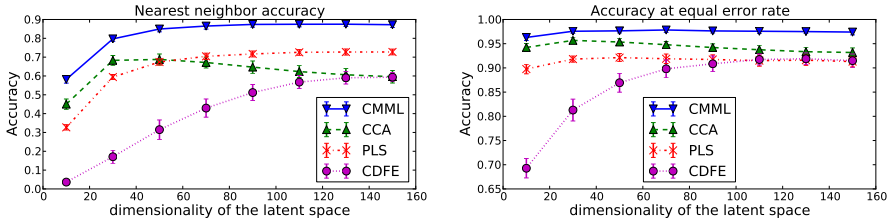

Fig. 6. CUFSF: nearest neighbor ccuracy (left) and equal error rate accuracy (right) for CMML ($n^- = 1$), CCA, PLS and CDFE.

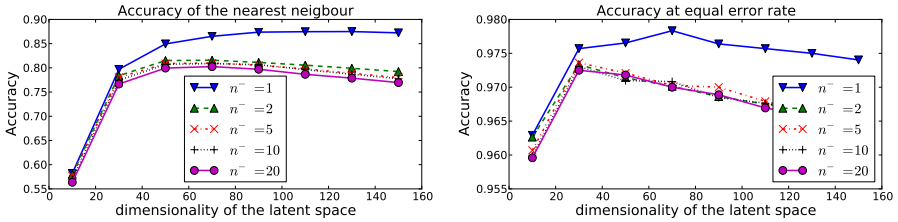

Fig. 7. CUFSF: nearest neighbor accuracy (left) and accuracy at equal error rate (right) using CMML for different ratios of negative/positive pairs n^-.

for the range of dimensionality studied. PLS follows similar trends as CMML but with an important gap in favor of CMML. In contrast with the results on Multi PIE, the performance of CCA on CUFSF, after a slight increase up to 30 dimensions, decreases for higher dimensions. CDFE follows the same trends as on Multi PIE, the performances consistently increase to saturate at higher dimensionality, reaching performance similar to CCA.

The behavior of CMML with the number of negative constraints on CUFSF (Fig. 7) is similar to the one observed on Multi PIE. The best results are obtained for one negative pair per positive pair, and the drop of performance with higher number of negative pairs tends to increase with the dimensionality of the latent space.

Results on LFW. In contrast with Multi PIE and CUFSF where the faces are taken in strictly controlled conditions of pose, lighting and expressions, LFW faces were captured "in the wild", exhibiting a wide range of variations. On this database, PLS, CCA and CDFE achieve poor performances. Performance of CDFE remains steady with the dimensionality of the latent space, but the accuracy is only slightly better than chance (roughly between 55 and 60%). PLS's accuracy is around 63% for 20 to 40-d latent space. CCA's performance steadily increases with the number of dimensions but at 150 dimensions its accuracy remains under 70%. In contrast, CMML reaches maximal accuracy with about 70 dimensions and then remains quite stable around 80%.

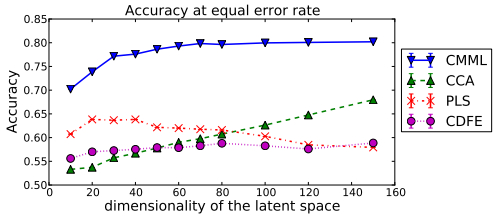

Fig. 8. Accuracy on LFW at equal error rate for CMML ($n^- = 1$), CCA, PLS and CDFE.

5 Conclusion

This paper addresses the problem of Cross Modal Matching by introducing CMML, a new metric learning method that learns discriminatively a low dimensional latent space in which object represented in different modalities can be directly compared. Experiments on three different databases for cross-modal face recognition showed that CMML is able to discover relevant low dimensional structures while remaining robust to over-fitting when the dimensionality of the latent space increases. Overall, CMML consistently outperforms other popular methods such as CCA, PLS and CDFE.

Acknowledgments:

This work was partly realized as part of the QUAERO Program funded by OSEO, French State agency for innovation and by the ANR, grant reference ANR-08-SECU-008-01/SCARFACE.

References

1. Hotelling, H.: Relations Between Two Sets of Variates. Biometrika **28** (1936) 321–377
2. Wold, H. Quantitative Sociology: International Perspectives on Mathematical and Statistical Modeling. In: Path models with latent variables: the NIPALS approach. Academic press edn. Academic Press, London (1975) 307–357
3. Lin, D., Tang, X.: Inter-modality face recognition. In: ECCV (4). (2006) 13–26
4. Li, Z., Lin, D., Meng, H.M., Tang, X.: Discriminant mutual subspace learning for indoor and outdoor face recognition. In: CVPR. (2007)
5. Wang, X., Tang, X.: Face photo-sketch synthesis and recognition. IEEE Trans. Pattern Anal. Mach. Intell. **31** (2009) 1955–1967
6. Gross, R., Matthews, I., Cohn, J.F., Kanade, T., Baker, S.: Multi-pie. Image Vision Comput. **28** (2010) 807–813
7. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts (2007)
8. Chen, J., Yi, D., Yang, J., Zhao, G., Li, S.Z., Pietikäinen, M.: Learning mappings for face synthesis from near infrared to visual light images. In: CVPR. (2009) 156–163

9. Liu, Q., Tang, X., Jin, H., Lu, H., Ma, S.: A nonlinear approach for face sketch synthesis and recognition. In: CVPR (1). (2005) 1005–1010
10. Goswami, D., Chan, C.H., Windridge, D., Kittler, J.: Evaluation of face recognition system in heterogeneous environments (visible vs nir). In: ICCV Workshops. (2011) 2160–2167
11. Klare, B., Li, Z., Jain, A.K.: Matching forensic sketches to mug shot photos. IEEE Trans. Pattern Anal. Mach. Intell. **33** (2011) 639–646
12. Borga, M., Landelius, T., Knutsson, H.: A unified approach to pca, pls, mlr and cca (1992)
13. Sharma, A., Jacobs, D.: Bypassing Synthesis: PLS for Face Recognition with Pose, Low-Resolution and Sketch. In: CVPR. (2011) 593–600
14. Li, A., Shan, S., Chen, X., Gao, W.: Face recognition based on non-corresponding region matching. In: ICCV. (2011) 1060–1067
15. Zhang, W., Wang, X., Tang, X.: Coupled Information-Theoretic Encoding for Face Photo-Sketch Recognition. In: CVPR. (2011)
16. Yi, D., Liu, R., Chu, R., Lei, Z., Li, S.Z.: Face matching between near infrared and visible light images. In: ICB. (2007) 523–530
17. Lei, Z., Li, S.Z.: Coupled spectral regression for matching heterogeneous faces. In: CVPR. (2009) 1123–1128
18. Zhou, C., Zhang, Z., Yi, D., Lei, Z., Li, S.Z.: Low-resolution face recognition via simultaneous discriminant analysis. Biometrics, International Joint Conference on **0** (2011) 1–6
19. Goldberger, J., Roweis, S., Hinton, G., Salakhutdinov, R.: Neighbourhood components analysis. In: Advances in Neural Information Processing Systems 17, MIT Press (2004) 513–520
20. Torresani, L., Lee, K.C.: Large margin component analysis. In Schölkopf, B., Platt, J., Hoffman, T., eds.: Advances in Neural Information Processing Systems 19. MIT Press, Cambridge, MA (2007) 1385–1392
21. Mignon, A., Jurie, F.: PCCA: A new approach for learning distances from sparse pairwise constraints. In: CVPR. (2012)
22. Journée, M., Bach, F., Absil, P.A., Sepulchre, R.: Low-rank optimization on the cone of positive semidefinite matrices. SIAM Journal on Optimization (**20**) 2327–2351
23. Burer, S., Monteiro, R.D.: A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. Mathematical Programming (series B **95** (2001) 2003
24. Chapelle, O.: Training a support vector machine in the primal. Neural Computation **19** (2007) 1155–1178
25. Tan, X., Triggs, B.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE Transactions on Image Processing **19** (2010) 1635–1650
26. Phillips, J.P., Moon, H., Rizvi, S.A., Rauss, P.J.: The FERET Evaluation Methodology for Face-Recognition Algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (2000) 1090–1104
27. Huang, G., Jones, M., Learned Miller, E.: LFW results using a combined Nowak plus MERL recognizer. In: Faces in Real-Life Images Workshop in European Conference on Computer Vision (ECCV). (2008)
28. Guillaumin, M., Verbeek, J., Schmid, C.: Is that you? metric learning approaches for face identification. In: International Conference on Computer Vision. (2009)
29. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. In: International Journal of Computer Vision. (1999) 1150–1157