An Overview and Empirical Comparison of Distance Metric Learning Methods

Panagiotis Moutafis, *Student Member, IEEE,* Mengjun Leng, *Student Member, IEEE,*
Ioannis A. Kakadiaris, *Senior Member, IEEE*

***Abstract*—In this paper, we first offer an overview of advances in the field of distance metric learning. Then, we empirically compare selected methods using a common experimental protocol. The number of distance metric learning algorithms proposed keeps growing due to their effectiveness and wide application. However, existing surveys are either outdated or they focus only on a few methods. As a result, there is an increasing need to summarize the obtained knowledge in a concise, yet informative manner. Moreover, existing surveys do not conduct comprehensive experimental comparisons. On the other hand, individual distance metric learning papers compare the performance of the proposed approach with only a few related methods and under different settings. This highlights the need for an experimental evaluation using a common and challenging protocol. To this end, we conduct face verification experiments, as this task poses significant challenges due to varying conditions during data acquisition. In addition, face verification is a natural application for distance metric learning because the encountered challenge is to define a distance function that: (i) accurately expresses the notion of similarity for verification, (ii) is robust to noisy data, (iii) generalizes well to unseen subjects, and (iv) scales well with the dimensionality and number of training samples. In particular, we utilize well-tested features to assess the performance of selected methods following the experimental protocol of the state-of-the-art database Labeled Faces in the Wild. A summary of the results is presented along with a discussion of the insights obtained and lessons learned by employing the corresponding algorithms.**

***Index Terms*—Face Recognition, Face Verification, Distance Metric Learning, Similarity Learning, Dimensionality Reduction**

I. Introduction

DINSTANCE metric learning has been an active area of research for many years, motivated by the need to properly define dissimilarity (or equivalently similarity) measures. An illustrative example is provided in Fig. 1. As demonstrated, there are different notions of similarity for face recognition applications (e.g., gender, pose, race, identity). Distance metric learning approaches are thus used to define a suitable distance metric that reflects what is considered to be "similar" or "dissimilar" in each case. In this paper, we use face verification as a case study because it poses significant challenges. Specifically, face verification is the problem of comparing two face images and determining whether or not they depict the same subject. Face data are characterized by large intra- and inter-personal variations due to differences in various attributes such as age, illumination, expression, and ethnicity. As a result, the features used as input to the matchers are usually of low quality for the purpose of face verification. Distance metric learning approaches offer a natural solution to this problem as: (i) they can be trained to accurately reflect the notion of similarity for the task at hand; (ii) they are robust to noisy data; (iii) they have the potential to generalize well to unseen classes without any re-training or adaptation; (iv) they offer capabilities of dimensionality reduction; (v) they can be combined naturally with non-linear classifiers such as the Nearest Neighbor (kNN) rule [1]; (vi) they are suitable for multi-class problems; (vii) they can accommodate problems with few samples per class; (viii) they can be used effectively for datasets with multi-modal classes; and (ix) they can be trained using weak types of constraints. Traditional classification methods such as Support Vector Machines (SVMs), Neural Networks (NNs), and Decision Trees (DTs) are limited in at least one of these aspects.

Fig. 1. Depiction of face images from the Labeled Faces in the Wild database [2]. As demonstrated, the notion of similarity varies (e.g., gender, pose, race, identity).

Due to the aforementioned advantages, distance metric learning approaches have been receiving increasing attention over the past few years. An illustration of this trend is depicted in Fig. 2. Yang [3] offered a comprehensive literature review to summarize the advances in this field. The methods in that technical report were organized in four groups: (i) supervised, (ii) unsupervised, (iii) SVM-based, and (iv) kernel-based. One year later, Yang [4] offered an updated summarization of the supervised and unsupervised categories to include the most recent papers. The papers in these two reviews are now outdated, as the most recent reference dates back to 2007. In another taxonomy, Ramanan and Baker [5] focused on local distance functions that can be considered approximations of

P. Moutafis, M. Leng, and I.A. Kakadiaris are with the Computational Biomedicine Laboratory (CBL), Department of Computer Science, University of Houston, 4800 Calhoun, Houston, TX 77204. (E-mail: {pmoutafis, mleng2, ioannisk@uh.edu}).

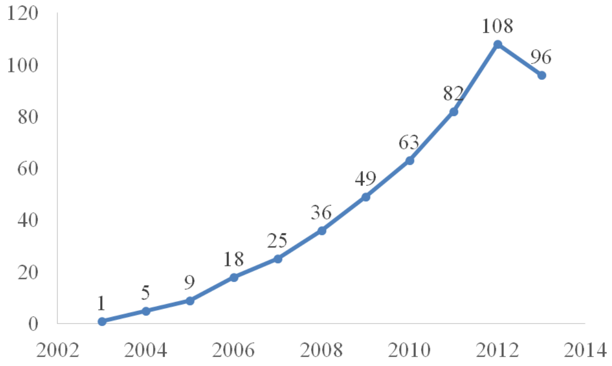

Fig. 2. Depiction of the number of papers published during the years 2003-2013 that include "Metric Learning" in their title according to the search engine Google Scholar.

geodesic distances defined by a metric tensor. They categorized these approaches according to three criteria; *how*, *where*, and *when* the distance metrics estimate the metric tensor. Even though their work provides valuable insights accompanied by interesting experimental results, its breadth is very limited. In particular, it focuses on only a few distance metric learning methods. The most recent is the journal version of the Large Margin Nearest Neighbor method published in 2009 [6]. Kulis [7] proposed a unified framework that can be used to describe large classes of metric learning approaches. Even though this paper offers an analysis of various aspects of distance metric learning and is a great introduction for new researchers in this area, it does not provide a comprehensive overview of the literature. In particular, the most recent reference that directly relates to distance metric learning dates back to 2011. Finally, Bellet et al. [8] offered a survey that compares related methods according to various criteria (e.g., scalability, form of metric, etc.). A distinguishing feature of this technical report is that it reviews recent trends and extensions of distance metric learning approaches to domains such as semi-supervised, multi-task, and structured data metric learning. However, it neglects many of the recent advances for the traditional task of single-domain supervised learning. Moreover, the conclusions drawn rely on evidence provided by the original papers without any empirical validation. In summary, the existing surveys are outdated [3], [4], too narrow [5], or they do not provide experimental evidence [8], [7].

To address these limitations, we first offer an overview of recent advances in the context of distance metric learning. Then, we assess the performance of selected methods using a common experimental protocol. For the first part, we conducted a systematic search of the literature that covered the years $2011-2013$. To make our search tractable we restricted ourselves to selected conferences and journals. Even though we are familiar with related papers published in other venues, a selection had to be made. The venues and keywords used are listed in Table I. Papers that include at least one of the keywords in their title were reviewed in more detail to determine their relevance. The venues were selected based on our experience in the field and considering the case study at hand (i.e., face recognition). To narrow down the scope of our search even further, we focused on methods that seek a single distance metric for supervised classification. Our objective is not to provide an in-depth analysis of each method nor to define a unified framework. Instead, we focus on covering as many papers as possible, since interested readers can read the corresponding papers themselves. That is, this paper should be viewed as complementary to the existing literature. In addition, we propose a taxonomy that classifies each method according to its main contribution or distinguishing feature. The proposed taxonomy comprises five categories: (i) ensemble, (ii) non-linear, (iii) regularized, (iv) probabilistic, and (v) cost-variant. In particular, a brief description is provided for each category and each approach. This contribution can assist researchers to place their work in the correct context within the literature and thus compare to appropriate methods. Moreover, a comparative summary of the methods under consideration is offered according to various criteria including dimensionality reduction and type of constraints used. Finally, we refer the reader to notable contributions that relate to distance metric learning but fall outside the scope of this paper. As mentioned, the existing literature lacks a comprehensive empirical evaluation of distance metric learning methods. To highlight this problem, we summarize the results reported in the literature for the state-of-the-art dataset *Labeled Faces in the Wild* (LFW) [2]. As demonstrated, several results are reported under different settings and different performance measures. This motivates the need for a comparative study that uses a common experimental protocol. In the second part of our paper, we address this problem by utilizing publicly available code to perform face recognition experiments using well-known, publicly available features for LFW. We follow the LFW protocols to directly compare the performance of selected methods. Furthermore, we investigate the effect of different factors such as feature length, type of constraints, and generalization capabilities. In addition to presenting and analyzing the obtained results, we also offer a discussion with the insights we gained by employing the various distance metric learning algorithms.

The rest of the paper is organized as follows: in Sec. II we discuss preliminaries; in Sec. III we present the proposed taxonomy and discuss key aspects of each category and each method; in Sec. IV we present and analyze the experimental results; and in Sec. V we offer a summary of our findings.

II. Preliminaries

In this section, we review basic concepts for readers who are not familiar with the distance metric learning domain. The advantages over other classification methods are also detailed. However, this section does not offer a thorough analysis of the field. For an in-depth understanding we refer interested readers to [7]. In addition to offering a brief description of main principles of distance metric learning, this section also serves as a guide for the terminology to be used throughout the paper. In particular, each sample is denoted by an n-th dimensional vector $\boldsymbol{x}_i \epsilon \mathbb{R}^n$, where i corresponds to the sample index. Given two samples $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, the Euclidean

TABLE I

THIS TABLE LISTS THE CONFERENCES AND JOURNALS USED FOR IDENTIFYING DISTANCE METRIC LEARNING METHODS. PAPERS THAT INCLUDE AT LEAST ONE OF THE KEYWORDS IN THEIR TITLE WERE CONSIDERED IN OUR SEARCH.

Conferences	Journals
AAAI Conference on Artificial Intelligence (AAAI)	Journal of Machine Learning Research (JMLR)
European Conference on Computer Vision (ECCV)	Artificial Intelligence (AI)
IAPR Int. Conference on Biometrics (ICB)	IEEE Trans. on Cybernetics (TCYB)
IEEE Conference on Automatic Face and Gesture Recognition (FG)	IEEE Trans. on Information Forensics and Security (TIFS)
IEEE Conference on Computer Vision and Pattern Recognition (CVPR)	IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)
Int. Conference on Biometrics: Theory, Applications and Systems (BTAS)	Machine Learning (ML)
Int. Conference on Computer Vision (ICCV)	Journal of Artificial Intelligence (JAIR)
Int. Conference on Machine Learning (ICML)	Pattern Recognition (PR)
Int. Joint Conference on Biometrics (IJCB)	
Neural Information Processing Systems Conference (NIPS)	
Keywords: Distance, Metric, Similarity, Mahalanobis, Embedding, Nearest, Neighbor, Face, Recognition	

distance is defined as $d_E(\boldsymbol{x_i},\boldsymbol{x_j})=\sqrt{(\boldsymbol{x}_i-\boldsymbol{x}_j)^\top(\boldsymbol{x}_i-\boldsymbol{x}_j)}$ or equivalently as the ℓ_2-norm, $d_E(\boldsymbol{x_i},\boldsymbol{x_j})=||\boldsymbol{x}_i-\boldsymbol{x}_j||_2$. This formula considers the square root of the inner product of the difference of two vectors. As a result, it is sensitive to the scaling and dimensionality of the features. More importantly, it does not utilize side-information and therefore it cannot accurately reflect what is considered to be similar for the task at hand. To address these limitations, most distance metric learning methods propose different ways of learning a Mahalanobis distance [30], which is defined as follows:

$$d_M(\boldsymbol{x}_i-\boldsymbol{x}_j)=\sqrt{(\boldsymbol{x}_i-\boldsymbol{x}_j)^\top\boldsymbol{A}^{-1}(\boldsymbol{x}_i-\boldsymbol{x}_j)}\ . \tag{1}$$

The matrix $\boldsymbol{A}\epsilon\mathbb{R}^{n\times n}$ in this formula is the Mahalanobis matrix that parameterizes the distance. It scales the features and utilizes their correlations to compute distances for ellipsoidally distributed data more effectively. When $\boldsymbol{A}$ is the identity matrix d_M reduces to d_E. Distance metric learning approaches learn $\boldsymbol{A}$ with the goal of minimizing a cost function subject to constraints defined by the data. The constraints can be represented in the form of: (i) labels, (ii) pairwise relationships, and (iii) proximity relation triplets. The labels can be used to derive the pairwise relationships, and the pairwise relationships to derive proximity relation triplets. In that sense, labels are the strongest form of constraints representation and proximity relation triplets the weakest. However, most distance metric learning algorithms work either with pairwise relationships or with proximity relation triplets derived from labels. There are a few methods that use quadruplet constraints (e.g., [24], [31]). Since this is not very usual we omit this part and refer interested readers to the corresponding papers. The labels are usually given in the form of $y_i\epsilon\{1,\cdots,c\}$ where y_i denotes the class label and c the number of classes. Pairwise relationships are defined in terms of pairs of samples that should be similar or dissimilar:

$$S=\{(\boldsymbol{x}_i,\boldsymbol{x}_j):\boldsymbol{x}_i \text{ and } \boldsymbol{x}_j \text{ should be similar}\}\ ,$$
$$D=\{(\boldsymbol{x}_i,\boldsymbol{x}_l):\boldsymbol{x}_i \text{ and } \boldsymbol{x}_l \text{ should be dissimilar}\}\ .$$

Proximity relation triplets are relative in the sense that specify the points that should be more similar than others:

$$R=\{(\boldsymbol{x}_i,\boldsymbol{x}_j,\boldsymbol{x}_l):\boldsymbol{x}_i \text{ more similar to } \boldsymbol{x}_j \text{ than } \boldsymbol{x}_l\}\ .$$

Most methods share the same goal expressed in different ways. That is, most approaches seek a distance that brings similar samples (samples of the same class) “closer”, while it “pushes away” dissimilar ones (samples of different classes). This problem, though, is formulated in various ways using different objective functions. To ensure that d_M satisfies the properties of a distance, $\boldsymbol{A}$ is usually restricted to be symmetric positive semi-definite. In some cases this restriction is relaxed to non-negativity and d_M reduces to a pseudo-metric. In other cases this restriction is completely dropped. To make computations easier the squared d_M formula is sometimes used instead:

$$d_M^2(\boldsymbol{x}_i-\boldsymbol{x}_j)=(\boldsymbol{x}_i-\boldsymbol{x}_j)^\top\boldsymbol{A}^{-1}(\boldsymbol{x}_i-\boldsymbol{x}_j)\ . \tag{2}$$

Since d_M^2 does not satisfy the triangle inequality, it is neither a metric nor a pseudo-metric. Some methods decompose the Mahalanobis matrix to $\boldsymbol{A}=\boldsymbol{L}^\top\boldsymbol{L}$. In this case the distance is defined as:

$$d_M(\boldsymbol{x}_i-\boldsymbol{x}_j)=||\boldsymbol{L}(\boldsymbol{x}_i-\boldsymbol{x}_j)||_2\ . \tag{3}$$

That is, the data are projected to another space by $\boldsymbol{L}$ and the Euclidean distance is computed using the projected data. By learning $\boldsymbol{L}\ \epsilon\mathbb{R}^{n\times d}$, where $d<n$ the data are projected to a lower-dimensional space $\mathbb{R}^d$. This strategy performs dimensionality reduction but breaks the convexity of Eq. (1). Nevertheless, computational and storage efficiency is obtained [8], [13], and the optimization problem usually converges stably to a local minimum [13], [32]. Finally, it has been observed that learning a similarity measure $s(\boldsymbol{x}_i-\boldsymbol{x}_j)=\boldsymbol{x}_i\boldsymbol{A}\boldsymbol{x}_j$ can sometimes be more efficient and more effective [33].

Distance metric learning methods offer a natural solution to the problem of face verification. By utilizing labels or pairwise relationships, such methods can learn a distance function that yields “small” values for pairs of images obtained from the same subject and “larger” values for pairs of images obtained from different subjects. That is, they can accurately reflect

TABLE II

PAPERS INCLUDED IN THE TAXONOMY. THE CONSTRAINTS COLUMN REFLECTS THE TYPE OF CONSTRAINTS UTILIZED ON THE OBJECTIVE FUNCTION OF THE RESPECTIVE METHODS. THE POSSIBLE ENTRIES FOR THE PROJECTION COLUMN ARE: (I) LINEAR, (II) NON-LINEAR, AND (III) KERNEL. THE NON-LINEAR TERM DENOTES METHODS THAT DIRECTLY OPTIMIZE THE OBJECTIVE FUNCTION IN A NON-LINEAR SPACE, WHILE THE KERNEL TERM DENOTES PAPERS THAT PRESENT KERNELIZED VERSIONS OF THE CORRESPONDING METHODS. THE OPTIMUM COLUMN CAN BE GLOBAL OR LOCAL DEPENDING ON THE CONVEXITY OR NON-CONVEXITY OF THE OBJECTIVE FUNCTION, RESPECTIVELY. THE COLUMN DIM. RED. DENOTES THAT THE CORRESPONDING PAPER CLEARLY INDICATES THAT THE ALGORITHM DEVELOPED CAN REDUCE THE DIMENSIONALITY OF THE INPUT. THE ONLINE CODE COLUMN INDICATES WHETHER SOURCE CODE IS PROVIDED ON THE WEBSITE OF AT LEAST ONE OF THE AUTHORS.

Category	Name	Constraints	Projection	Optimum	Dim. Red.	Online Code
Ensemble	BoostMDM [9]	Proximity	Linear	Local	Yes	No
	BoostMetric [10]	Proximity	Linear	Global	No	Yes
	MetriBoost [11]	Proximity	Linear	Global	No	No
	REMetric [12]	Pairwise	Linear	Local	Yes	No
Non-linear	GB-LMNN [13]	Proximity	Non-linear	Local	Yes	Yes
	HDML [14]	Pairwise & Proximity	Non-linear	Local	Yes	Yes
	IPLR [15]	Pairwise	Non-linear	Global	Yes	No
	SLLC [16]	Proximity	Non-linear	Global	Yes	No
Regularized	DRMetric [17]	Proximity	Linear	Local	No	No
	R-MLR [18]	Proximity	Kernel	Global	No	Yes
	SRML [19]	Varies	Varies	Varies	Varies	No
	Sub-SML [20]	Pairwise	Linear	Global	No	Yes
Probabilistic	KISSME [21]	Pairwise	Linear	Global	No	Yes
	NCMML [22]	Labels	Linear	Local	Yes	Yes
	PPCA [23]	Pairwise	Kernel	Global	No	No
	RDC [24]	Quadruplets	Linear	Local	Yes	Yes
Cost-Variant	DML-eig [25]	Pairwise	Linear	Local	No	Yes
	DMPL [26]	Proximity	Linear	Local	No	No
	MELM [27]	Pairwise	Linear	Global	No	No
	aSMM [28]	Pairwise	Linear	Global	Yes	No
	NRML [29]	Proximity	Linear	Local	No	No

the notion of similarity for the task at hand. As a result, they offer numerous advantages over traditional classification methods. In particular, some of the most popular classification methods (e.g., SVMs, NNs, DTs) rely only on the labels. Their objective is to utilize statistical properties in the data to model each class individually. As a result, they require many samples per class and they are not as effective for problems with multi-modal classes. Moreover, they can classify single samples but they cannot measure the degree of similarity when a pair of samples is given as input. This also implies that they need to be retrained or adapted in order to be able to generalize to unseen classes. Hence, they do not scale well with the number of classes. Distance metric learning methods address all of these limitations, which are very important for the task of face verification. Two other challenges of face verification addressed by distance metric learning algorithms are the high dimensionality of the features and the noisy nature of the data due to the varying conditions during data acquisition. To address the first challenge, such methods learn a projection to a lower-dimensional space. For the latter, they incorporate appropriate regularization terms in their objective function. Finally, they can naturally be coupled with other classifiers (e.g., SVMs, kNN) for different kinds of problems.

III. TAXONOMY

In this section, we first offer an overview of selected papers published during the years 2011-2013. Then, we refer the reader to related approaches that we came across in our search. Most papers focus on a single challenge (e.g., intra-class variations) and adopt a principal approach to address it. We propose a clustering of the presented papers according their most prominent feature to: (i) help the reader focus on the main novelty introduced by each paper; and (ii) provide some form of structure of the existing literature. Specifically, we propose a taxonomy of selected papers in five categories: (i) *ensemble* approaches that learn many weak distance metrics (similar to weak classifiers), which are then combined into a single metric distance; (ii) *non-linear* methods that optimize the objective function directly in a non-linear space; (iii) *regularized* techniques that include a regularization term to meet the learning objectives; (iv) *probabilistic* models that optimize likelihood criteria; and (v) *cost-variant* algorithms that propose improved cost functions for the task at hand. Papers that fit in more than one category are classified according to their most distinguishing feature according to our understanding. To select the papers included in our paper we employed criteria such as applicability to the case study at hand and relevance according to our own subjective judgment. Due to space constraints we omit part of the literature. A summary of different attributes for the approaches in this survey is provided in Table II. From the papers included in our overview, nine are not cited by the existing surveys, while 13 are only cited but not discussed, and 17 are treated superficially.

Ensemble: Approaches in this category attempt to convey the benefits of ensemble-based approaches such as Boosting [34], Bootstrap Aggregating (bagging) [35], and AdaBoost [36] in the domain of distance metric learning. By combining many weak learners, increased robustness to noise and reduced time complexity can be achieved. Chang [9] proposed a *Boosting Mahalanobis Distance Metric* (BoostMDM) method. It iteratively employs a base-learner to update a base matrix. A framework to combine base matrices is developed in this paper, along with a base learner algorithm specific to it. The cost function minimizes the hypothesis margin, which is a lower bound of the sample margin used in methods such as SVMs. Since it is computed using the nearest hit and nearest miss of each sample it implicitly relies on proximity relation triples, which are updated on each iteration. Two extensions are proposed (i.e., BOOSTMDM-K and BOOSTMDM-G) to increase the stability of the algorithm and a regularization is introduced to keep the rank of the Mahalanobis matrix as small as possible. Convergence and an error bound are guaranteed only under certain assumptions. Shen *et al.* [10] proposed a *Boosting-Based Metric Learning* (BoostMetric) method. Their approach is based on the observation that any positive semi-definite matrix can be decomposed into a linear combination of trace-one rank-one matrices. Thus, many weak learners are learned and combined. The different algorithms developed in this paper rely on hinge loss, exponential loss, and logistic loss functions. To enhance performance, a multi-pass learning approach is investigated that updates the proximity relation triplets according to the projected data. A theorem is shown that demonstrates global convergence in the limit. Bi *et al.* [11] adapted the AdaBoost algorithm for Mahalanobis distance metric (MetriBoost). It combines many rank-one positive semi-definite matrices (i.e., weak learners) to learn a Mahalanobis matrix. Even though the algorithm is developed using proximity relation triplets, a bipartite strategy is employed to decompose the proximity relation triplets into pairwise relationships. The framework presented can be implemented using different weak metrics and weight parameters. Kozakaya *et al.* [12] proposed a *Random Ensemble Metric Learning* (REMetric) method. The proposed approach iteratively sub-samples the training data to extract two groups of samples using label information and learns a corresponding linear SVM each time. The grouping of the data can also be determined using pairwise relationships. The projections learned are integrated to form a Mahalanobis matrix. This approach can perform dimensionality reduction by reducing the number of projection vectors used to define the Mahalanobis matrix. As demonstrated, the projection matrix has the property of decorrelating the features.

Non-linear: The objective of non-linear methods is to learn more flexible metrics that can fit the data in a better way. Kedem *et al.* [13] proposed two non-linear extensions of the Large Margin Nearest Neighbor (LMNN) method [13] (i.e., χ^2-LMNN, and Gradient-Boosting LMNN). The χ^2-LMNN learns a linear mapping but optimizes the objective function according to the non-linear χ^2 distance. The Gradient-Boosting LMNN (GB-LMNN) learns a non-linear mapping directly in the function space by employing gradient-boosted regression trees. Both methods can perform dimensionality reduction. Norouzi *et al.* [14] proposed a *Hamming Distance Metric Learning* (HDML) approach. The proposed framework learns a discrete mapping of the input to binary codes with discriminative properties. It can be employed in conjunction with different families of hash functions and two types of losses: (i) pairwise hinge, and (ii) triplet ranking. Its main advantage is that the binary codes are storage efficient and allow for sub-linear k-NN search. The non-convexity of the method is mitigated by constructing a continuous upper bound on the empirical loss. Jain *et al.* [15] established a theoretical connection between *Metric and Kernel Learning*. They show that for the class of spectral functions, learning an optimal kernel (metric) can be used to compute the corresponding optimal metric (kernel). Therefore, many distance metric learning techniques can be efficiently kernelized. They also show that this result can be used to compute distances of unseen samples to the kernel space, alleviating the need to re-train the algorithm. However, the LogDet algorithm that they propose scales badly with the number of parameters. To address this problem, a parameter reduction approach is intorduced, which is called high-dimensional Identity Plus Low-rank (IPLR) metric learning. Bellet *et al.* [16] proposed a *Similarity Learning for Linear Classification* (SLLC) approach. The authors exploit results on *good similarities* [37] to seek a similarity function optimized in a non-linear space. The matrix learned is not required to be positive semi-definite, which reduces the time complexity of the proposed solution. The function learned is used to define a global linear classifier. An upper bound for the generalization error is shown based on the method's uniform stability. This approach satisfies all the necessary requirements needed to invoke the Kernel PCA trick [38]. As shown in the paper, the generalization bound is independent of the size of the projection space.

Regularized: By using appropriate regularization terms, distance metric learning approaches have the potential to become more robust to noise and generalize better. Liu *et al.* [17] proposed a *Doubly Regularized Metric* (DRMetric) approach. Their objective is to increase the efficacy of the learned weak distance metrics for boosting-based approaches. To this end, two regularization terms are used: (i) a total Kullbak-Leibler divergence regularization is applied on the weight of the training examples to smooth the changes of the weights and address the negative effect of outliers; (ii) a second regularization is applied to the rank-one matrices to be combined, which makes them less redundant and reduces the number of rank-one matrices needed. Lim *et al.* [18] proposed a *Robust Metric Learning to Rank* approach (R-MLR) that extends the Metric Learning to rank [39] algorithm. The proposed extension enforces a group sparsity penalty on the learned transformation to promote sparsity of the input features. This results in detection and suppression of irrelevant features. At the same time a trace penalty is imposed to promote sparsity of the output projections. Consequently, the matrix learned is promoted to have low-rank, which simplifies the model complexity and improves robustness. Even though this scheme is based on a ranking approach it can be applied to classification tasks as well. Jiang *et al.* [19] proposed a *Sparsity-Regularized Metric Learning* (SRML) approach, motivated by the problem of determining the optimal dimensionality for the learned discriminative projection. To this end, a regularization is imposed that maximizes the number of all-zero rows in the projection matrix. These rows are then removed from the projection matrix. As illustrated, this regularization is applicable to different approaches. Besides benefits in terms of accuracy, reducing the dimensionality also reduces the number of parameters to be estimated. Cao *et al.* [20] proposed a novel regularization framework for *Similarity Metric Learning Over the Intra-Personal Subspace* (Sub-SML) approach. This method is optimized for the task of face verification. Its main objectives are robustness to noise from intra-personal variations and class-separation based on identity. The first objective is met by mapping the Eigenfaces [40] to the intra-personal subspace. The second objective is met by learning a generalized similarity metric, which is simultaneously parametrized by a Mahalanobis distance and a similarity metric.

Probabilistic: By approaching the problem from a probabilistic point of view, methods in this category can reduce the time complexity and avoid over-fitting the data. Köstinger *et al.* [21] proposed a *Keep it simple and straightforward metric learning* (KISSME) method. Their approach relies on pairwise relationships and considers two independent processes that generate similar and dissimilar points. By further assuming a Gaussian data distribution, a Mahalanobis distance is learned by optimizing a likelihood ratio test. This approach relies upon closed-form solutions derived by known maximum likelihood estimators and thus has a low computational cost. Mensink *et al.* [22] proposed a *Nearest Class Mean Metric Learning* (NCMML) approach. Unlike most methods that focus on NN classification, their algorithm learns a Mahanalobis distance suitable for Nearest Class Mean (NCM) [41] classification. In particular, the Mahalanobis matrix is learned by computing the maximum likelihood of a probabilistic model based on multiclass logistic regression. The key idea is to separate the classes that are nearby in the projected space to ensure correct predictions. In addition, a non-linear NCM extension is presented, which represents each class using multiple centroids. Mignon and Jurie [23] proposed the *Pairwise Constrained Component Analysis* (PCCA). The problem addressed is distance metric learning when only sparse pairwise relationships are available. The proposed algorithm uses a generalized logistic loss function to penalize distances of similar points that are larger than a given threshold and distances of dissimilar points that are smaller than the same threshold. The method can perform dimensionality reduction and a kernelized version is also presented. Zheng *et al.* [24] proposed the *Relative Distance Comparison Learning* (RDC) approach, motivated by the re-identification problem for which large variations exist within the samples of each class. To capture this information, while avoiding over-fitting issues, they propose a probabilistic approach that maximizes the probability of correct classifications. To increase the robustness for under-sampled training data, their model optimizes individual comparisons between any two data points.

Cost-Variant: Methods in this category propose new formulations of the cost function to meet different objectives. Ying and Li [25] proposed *Eigenvalue Metric Learning*. They presented a re-formulation of two popular methods. In the first case the objective of maximizing the sum of distances between dissimilar pairs is replaced by that of maximizing the minimal squared distances of dissimilar pairs. They call this approach DML-eig. In the second case a min-max formulation is presented that extends LMNN, thus named LMNN-eig. These two modifications allow for the development of first-order algorithms that depend on the computation of the largest eigenvector on each iteration. As a result, significant gains are obtained in terms of time complexity. Köstinger *et al.* [26] proposed a *Discriminative Metric and Prototype Learning* (DMPL) algorithm. The proposed method jointly learns the positioning of prototypes and a corresponding Mahalanobis distance metric to reduce the time complexity during testing. The obtained results indicate that not only a reduced time complexity is achieved, but the generalization performance is also improved. Li and Shan [27] proposed a *Margin Emphasized Metric Learning* (MEML) approach. They presented a new objective function that adds more weight to pairs of samples that are close to the decision boundary. To this end, points are re-weighted at each iteration. The choice of the re-weighting functions is empirically determined. Yu *et al.* [28] proposed a *Minimal Distance Maximization* approach. They seek to address the robustness issues of dimensionality reduction methods, such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) [42]. To this end, they proposed a minimal distance maximization algorithm that maximizes the minimal between-class distances in the projected space. Since this non-convex model makes parametric assumptions, the authors also developed a non-parametric extension and a convex approximation model called aSMM. Lu *et al.* [29] proposed a *Neighborhood Repulsed Metric Learning* (NRML) approach. Their approach seeks a distance metric that brings samples from the same class "closer", while it "pushes" samples from different classes as far as possible. The cost is computed locally and the learning process emphasizes more on the latter goal. A version of the method that exploits multiple feature descriptors is also developed (MNRML).

Finally, we cite recent advances for research problems within or closely related to the context of distance metric learning. The provided bibliography is not an exhaustive list of the literature. In the process of searching papers for our work, we came across approaches for *adaptive* [43], [44], *multi-task* [45], *online* [46], *semi-supervised* [47], [48], [49], [50], *unsupervised* [51], and *transfer* [52], [53] learning. In addition, we found many papers that take into consideration special forms of data including *time series* [54], *structured* [55], [56], [57], [58], [31], [59], *multi-label, multi-view and bags* [60], [61], [62], [63], [64], [65]. Other papers address the problem of *heterogeneous domains* [66], [67], *regression* [68], [69], *point to set classification* [70], [71], [72], [73], [74],

[75], and *manifold learning* [76], [77], [78]. Finally, *latent models* [79], [80], [81], [82], theoretical results [83], [84], and other papers related to kNN classification and distance metric learning [85], [86], [87], [88], [89], [90] have recently been published.

IV. Experimental Evaluation

In this section, we first utilize results from the literature to highlight the need for a broad empirical evaluation using a common experimental protocol. Then, we describe the dataset used in our experiments and discuss the corresponding implementation details. Finally, we present the obtained results.

In particular, we present a summary of the existing results found in the literature for LFW. This dataset was selected because it is the most frequently used among the papers listed in Table II. In these papers, several results are reported under different settings. In Table III and Fig. 3, we offer an overview of the results found in the literature when the same feature extraction method and the same experimental protocol were used. Nevertheless, as demonstrated by Table III, the dimensionality of the input features is still different because different post-processing methods are used in each case. In addition, some of the corresponding papers omit some performance measures (e.g., accuracy). Hence, the existing results are not comparable and a broad empirical evaluation using the same experimental protocol is needed. To address this limitation, we conducted experiments using methods from Table II that provide online MATLAB code.

A. Dataset

The Labeled Faces in the Wild dataset [2] was designed to facilitate research in unconstrained face recognition (i.e., face images acquired in uncontrolled environments). It comprises 13,223 face images captured from 5,749 different subjects. The number of images per subject ranges from one to 530. For 1,680 of the subjects, two or more images are available. LFW facilitates the comparison of different distance metrics for many reasons. First, images in this dataset were collected from news articles online. As a result, they exhibit natural variation in terms of focus, resolution, illumination, background, pose, facial expression, age, race, gender, accessories, makeup, and photographic quality. This makes LFW suitable to assess the robustness of different distance metrics. Second, the provided training-testing splits are mutually exclusive. That is, subjects in the testing set are not part of the training set. Hence, the generalization performance of the distance metric methods under consideration can be effectively assessed. Finally, the LFW database defines a detailed experimental protocol that researchers should follow. Therefore, the results obtained are directly comparable. Specifically, two views and two paradigms are provided. *View 1* consists of two subsets of images: (i) the training subset contains 1,100 similar pairs and 1,100 dissimilar pairs, and (ii) the testing subset contains 700 similar pairs and 700 dissimilar pairs. This view is not designed to evaluate the performance of the proposed methods. Instead, it should be used for algorithm development (e.g., parameters selection). *View 2*, on the other hand, is designed to assess the performance of the algorithms developed in *view 1*. This way, inappropriately fitting a distance metric to the test data is avoided. In particular, a ten-fold cross validation scheme is employed for *view 2*. That is, images are divided into ten folds, where 300 similar pairs and 300 dissimilar pairs are provided in each fold. The subjects are mutually exclusive across the folds. Each time, images that correspond to nine of the ten folds are used for training and images that correspond to the remaining one are used for testing. The unrestricted paradigm provides the subject's identity (ID) for each image. As a result, labels, pairwise relationships, and proximity relation triplets can easily be generated. On the contrary, the restricted paradigm does not allow researchers to access the IDs of the images. Instead, a set of similar and dissimilar pairs of images is provided. Algorithms that rely on labels and proximity relation triplets should infer this information utilizing the similar and dissimilar pairs.

TABLE III
Summary of results found in the literature for the restricted setting of LFW. The accuracy is in the format of mean (standard deviation). The dim. denotes the input feature dimensionality.

Method	dim.	Accuracy (%)			
		SIFT		Square Root of SIFT	
Sub-SML	300	85.55	$(6.1{\cdot}10^{-3})$	86.22	$(2.7\ {\cdot}10^{-3})$
DML-eig	100	80.55	$(0.2{\cdot}10^{-3})$	81.27	$(0.2{\cdot}10^{-3})$
REMetric	5000	75.60	$(0.1{\cdot}10^{-3})$	76.00	$(0.2{\cdot}10^{-3})$
KISSME	100	-	-	-	-
DRMetric	400	-	-	-	-
PCCA	-	83.80	$(4.0{\cdot}10^{-3})$	-	-

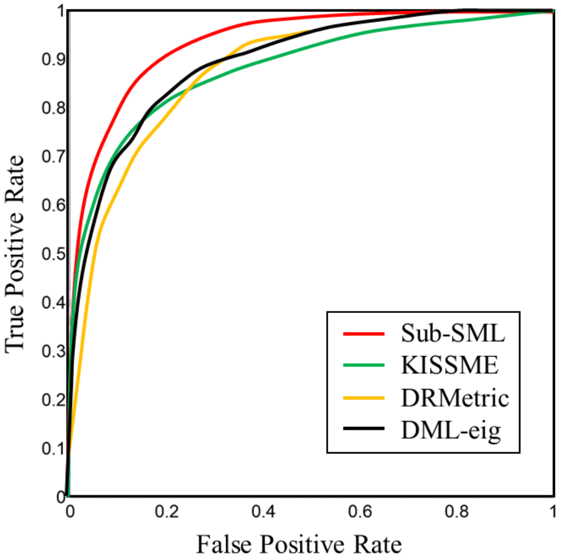

Fig. 3. Depiction of ROC curves for the restricted setting of LFW.

B. General Implementation Details

Distance metric learning algorithms usually require two inputs: features and constraints. The following steps ensure

TABLE IV
SUMMARY OF RESULTS FOR EXPERIMENT 1. THE VALUES OF ACCURACY (ACC), VERIFICATION RATE (VR), AREA UNDER CURVE (AUC), AND TRAINING TIME (T) ARE GIVEN IN THE FORMAT MEAN (STANDARD DEVIATION). THE TOP PART OF THE TABLE REFERS TO THE RESTRICTED PARADIGM, AND THE BOTTOM PART TO THE UNRESTRICTED ONE. THE ΔACC, ΔVR AND ΔAUC DENOTE THE RELATIVE CHANGES OBTAINED FOR EACH METHOD.

Method	Absolute Value					Relative Difference		
	ACC (%)	VR (%)	AUC (%)	T(s)	Dim.	ΔACC (%)	ΔVR (%)	ΔAUC (%)
Baseline	70.13 (1.15)	66.87 (1.45)	77.40 (1.34)	0 (0)	100 (0)	0.00	0.00	0.00
DML-eig	76.93 (1.32)	74.13 (3.04)	85.66 (1.12)	6 (2)	100 (0)	9.70	10.87	10.67
KISSME	81.55 (2.05)	78.50 (4.29)	89.91 (1.53)	**0 (0)**	100 (0)	16.28	17.40	16.17
Sub-SML	**84.43** (1.68)	**83.39** (1.97)	**92.56** (0.99)	39 (1)	100 (0)	**20.39**	**25.47**	**19.59**
RDC	73.68 (1.07)	76.73 (2.46)	81.82 (1.30)	5,015 (178)	**46** (1)	5.06	14.76	5.71
NCMML	76.77 (1.65)	73.40 (4.42)	85.58 (1.05)	41 (1)	100 (0)	9.46	9.77	10.57
BoostMetric	76.97 (1.86)	73.37 (4.00)	85.02 (1.13)	116 (6)	100 (0)	9.74	9.72	9.84
R-MLR	73.65 (1.73)	73.13 (3.68)	82.36 (1.26)	3,486 (730)	100 (0)	5.01	9.37	6.41
GB-LMNN	78.10 (1.91)	71.70 (4.54)	86.95 (1.79)	2,911 (97)	100 (0)	11.36	7.23	12.34
Baseline	70.32 (1.36)	65.07 (2.39)	77.40 (1.34)	0 (0)	100 (0)	0.00	0.00	0.00
DML-eig	77.45 (2.06)	75.07 (3.77)	86.17 (2.15)	4 (0)	100 (0)	10.14	15.37	11.34
KISSME	80.97 (2.34)	75.60 (4.37)	89.82 (1.60)	**0** (0)	100 (0)	15.15	16.19	16.50
Sub-SML	**83.57** (1.34)	**79.40** (2.40)	**92.33** (0.90)	82 (5)	100 (0)	**18.84**	**22.03**	**19.29**
RDC	74.28 (1.55)	75.77 (3.61)	82.29 (1.88)	17,610 (2,062)	**41** (7)	5.64	16.44	6.32
R-MLR	81.43 (1.88)	78.63 (4.03)	90.25 (1.20)	2,002 (344)	100 (0)	16.56	20.66	16.60
NCMML	78.72 (1.38)	75.57 (3.46)	87.14 (1.03)	43 (2)	100 (0)	12.67	15.96	12.58
BoostMetric	78.57(2.10)	75.37 (3.30)	86.78 (1.46)	205 (6)	100 (0)	12.45	15.65	12.13
GB-LMNN	77.80 (1.75)	67.57 (4.72)	86.62 (1.50)	9,520 (187)	100 (0)	11.35	3.68	11.92

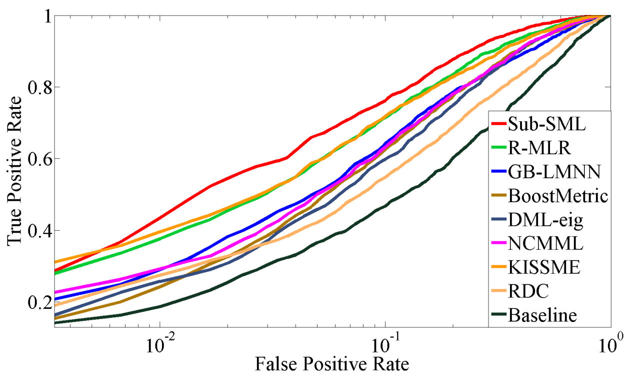

Fig. 4. ROC curves for the unrestricted setting of experiment 1.

that both have the appropriate form: (i) feature extraction, (ii) post-processing, and (iii) constraint generation. The feature extraction and post-processing steps were used to reduce the dimensionality of the features, while retaining and/or enhancing their discriminative properties. A constraint generation step was also performed to ensure that the constraints at hand meet the minimum requirements of the distance metric learning algorithms employed.

Feature Extraction: The Scale-Invariant Feature Transform (SIFT) features were used, as extracted by Guillaumin *et al.* [91] using the LFW funneled registered images [92]. These features have been used by many papers in the literature [91], [12], [25], [17], [93] and they have been found to have good discriminative properties. In particular, Guillaumin *et al.* [91] computed SIFT descriptors at nine fixed points of the face (i.e., corners of mouth, eyes, and nose), detected by a landmark detector [94]. For each point, 128 dimensional SIFT descriptors were computed at three different scales. Hence, each image is represented by a 9×128×3 = 3,456 dimensional vector. However, the performance that these features yield is degraded for many reasons. For example, the landmark detection is not always accurate. This results in noisy features as they encompass non-discriminative information. Moreover, they are high dimensional, which increases the computational time and space cost. In addition, the performance of these features is affected by large intra-subject variations such as pose and illumination conditions. Therefore, performing an appropriate post-processing step is necessary.

Post-processing: Principal Component Analysis (PCA) [95] is the most commonly used method to reduce the dimensionality of the features and reduce noise. Specifically, it computes the eigenvalue decomposition of the covariance matrix defined as $C = \sum_{x_i \in Tr} (x_i - \boldsymbol{\mu})(x_i - \boldsymbol{\mu})^T$, where Tr is the set of training feature vectors and $\boldsymbol{\mu}$ is the mean vector of all the training samples. The LFW protocol prohibits the use of the test set to estimate model parameters of any form. To this end, the features of the test set were projected to a low dimensional space by using the same principal components learned from the training set. In our experience, setting the dimensionality of the SIFT features between 100 and 500 provides a good trade-off between efficiency and effectiveness. This is in line with the findings of [20], [25]. To reduce large variations across different subjects, the norm of each feature vector was normalized to one.

Constraint Generation: The distance metric learning algorithms considered in our experiments rely either on labels or

TABLE V

SUMMARY OF RESULTS FOR EXPERIMENT 2. THE VALUES OF ACCURACY (ACC), VERIFICATION RATE (VR), AREA UNDER CURVE (AUC), AND TRAINING TIME (T) ARE GIVEN IN THE FORMAT MEAN ABSOLUTE VALUE (RELATIVE IMPROVEMENT). THE δACC, δVR AND δAUC DENOTE THE PERFORMANCE IMPROVEMENTS OBTAINED BY INCREASING THE FEATURE DIMENSIONALITY FROM 100 TO 300 AND 500, RESPECTIVELY.

Method	Feature Length: 300				Feature Length: 500			
	ACC (δACC)	VRv (δVR)	AUC (δAUC)	T (δT)	ACC (δACC)	VRv (δVR)	AUC (δAUC)	T (δT)
Baseline	71.07 (+1.07)	66.07 (+1.54)	78.45 (+1.90)	-	71.45 (+1.61)	66.07 (+1.54)	79.23 (+2.37)	-
DML-eig	79.55 (+2.71)	75.00 (-0.09)	88.41 (+2.59)	53 (+1231)	79.20 (+2.26)	74.10 (-1.29)	88.40 (+2.59)	255 (+6308)
KISSME	78.10 (-3.54)	64.23 (-15.04)	87.86 (-2.18)	**2** (+365)	72.32 (-10.68)	48.60 (-35.71)	85.00 (-5.36)	**2** (+485)
Sub-SML	**85.32** (+2.09)	**80.30** (+1.13)	**93.43** (+1.20)	372 (+354)	**83.60** (+0.04)	74.13 (-6.63)	**92.97** (+0.70)	316 (+285)
RDC	74.67 (+0.52)	76.20 (+0.57)	82.99 (+0.86)	17,613 (0)	74.72 (+0.58)	75.63 (-0.18)	83.21 (+1.12)	16,402 (-7)
R-MLR	82.68 (+1.53)	75.13 (-4.45)	91.53 (+1.42)	1,318 (**-34**)	82.32 (+1.08)	74.37 (-5.43)	91.55 (+1.44)	1844 (**-8**)
NCMML	79.45 (+0.92)	77.03 (+1.94)	88.10 (+1.11)	121 (+182)	79.65 (+1.19)	**77.17** (+2.12)	88.45 (+1.50)	253 (+492)
BoostMetric	80.22 (+2.10)	75.53 (+0.22)	89.06 (+2.63)	1,793 (+775)	80.52 (+2.48)	75.10 (-0.35)	89.51 (+3.14)	5,347 (+2,508)
GB-LMNN	81.37 (**+4.58**)	73.97 (**+9.47**)	90.60 (**+4.59**)	42,102 (+342)	82.67 (**+6.26**)	75.73 (**+12.09**)	91.16 (**+5.24**)	86,656 (+810)

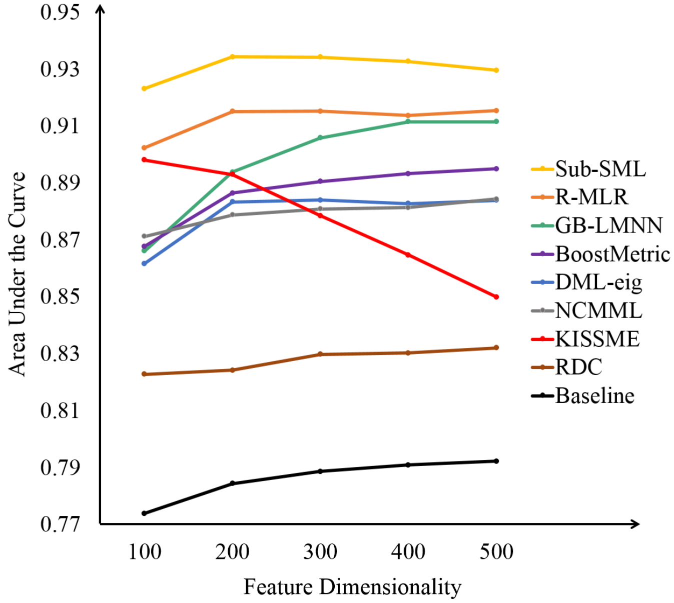

Fig. 5. Depiction of the mean AUC values obtained for input features with varying dimensionality obtained for experiment 2.

pairwise relationships. That is, even though the corresponding methods may sometimes accommodate weaker forms of constraints, the implementations provided by the authors use labels or pairs of similar/dissimilar samples. In the restricted paradigm, the pairwise relationships are directly provided by the dataset. To infer the label information the similar pairs were first parsed seeking relationships of hard transitivity. Samples that satisfied this relationship were assigned the same label. That is, if $(x_i, x_j)\epsilon\mathcal{S}$, and $(x_j, x_l)\epsilon\mathcal{S}$, then $y_i = y_j = y_l$. In the unconstrained paradigm, the labels are implicitly provided by the dataset in the form of IDs. Using this information it is possible to generate more pairwise relationships compared to the ones provided in the restricted paradigm. To generate similarity relationships, two images for each subject were randomly selected from the training set. That is, if $y_i = y_j$, then $(x_i, x_j)\epsilon\mathcal{S}$. To generate dissimilarity relationships, two images were randomly selected, each one from a different subject. That is, if $y_i \neq y_j$, then $(x_i, x_j)\epsilon\mathcal{D}$. Considering that imbalanced numbers of similar and dissimilar pairs may lead to unfair evaluation of the distance metric algorithms, the same number of similar and dissimilar pairs was generated. Restricted by the maximum number of similar pairs, 500, 600, and 700 similar and dissimilar pairs were generated for each fold, respectively. Finally, we provide general information concerning the training and testing parts of our experimental evaluation. Following the LFW protocol, nine of the ten folds were used for training the distance metric learning algorithms. The learned projections were used to project both training and testing data to the new space and the Euclidean distance between all pairs was computed. To assess the performance of the distance metric learning methods employed, the following measures are used: accuracy (ACC), verification rate (VR), area under curve (AUC) for the corresponding receiver operating characteristic curve (ROC), and computational training time (T). In particular, T denotes the time used to calculate the projection matrix. To assess the generalization properties of each method the corresponding training accuracy (ACC_T), training verification rate (VR_T) and area under curve in the training set (AUC_T) are recorded. The threshold that maximizes the accuracy on the training set was used to compute both the accuracy and verification rate values.

C. Experimental Results

Experiment 1: The objectives of this experiment are to assess the impact on the verification performance when selected distance metric learning methods are employed under the: (i) restricted paradigm, and (ii) unrestricted paradigm. For this experiment, the dimensionality of the input features was set to 100. In the restricted paradigm, the pairwise relationships were used to generate labels following the procedure described in Sec. IV-B. In the unrestricted paradigm, 500 similar pairs and 500 dissimilar pairs were generated for

TABLE VI
SUMMARY OF RESULTS FOR EXPERIMENT 3. THE σACC, σVR, AND σAUC DENOTE THE PERFORMANCE IMPROVEMENTS OBTAINED BY INCREASING THE NUMBER OF PAIRWISE CONSTRAINTS USED FOR TRAINING.

Method	similar/dissimilar pairs per fold: 600				similar/dissimilar pairs per fold: 700			
	σACC (%)	σVR (%)	σAUC (%)	σT (%)	σACC (%)	σVR (%)	σAUC (%)	σT (%)
Baseline	-0.43	0.15	0	-	-0.64	0.15	0	-
DML-eig	**1.66**	1.82	**0.73**	**-22**	**1.10**	1.42	**0.99**	**-24**
KISSME	1.05	**1.90**	0.20	-46	0.97	**2.82**	0.33	-39
Sub-SML	0.42	0.17	0.14	5	0.89	1.05	0.23	12
RDC	-0.76	-1.06	-0.12	-22	-0.20	-2.51	0.35	28

each fold. The obtained results are summarized in Table IV and partial results are depicted in Fig. 4. To meet the first objective, the top part of Table IV reports the absolute and relative performance obtained under the restricted paradigm. To meet the second objective, the bottom part of Table IV reports the corresponding results obtained for the unrestricted paradigm. In both cases, the relative performance is computed in relation to the baseline. For example, ΔACC is defined as $\frac{ACC_m - ACC_b}{ACC_b}$, where ACC_m is the mean accuracy achieved by distance metric learning method m, and ACC_b is the mean accuracy of the baseline. Herein, the term baseline is used to denote the performance obtained using the processed features as explained in Sec. IV-B. That is, the Euclidean distance is computed using the processed features without employing any distance metric learning method. The term *Dim.* is used to denote the dimensionality of the output feature space. Since the training time for HDML exceeds 60 hours, the corresponding results are not reported. In both paradigms, Sub-SML yields the best verification performance. In our experience, this is due to its intra-subspace projection in conjunction with the effectiveness of the regularization proposed by the authors. The verification performance of KISSME ranks second in most cases. However, it is much faster than the other methods. In particular, its average training time is only 0.3 seconds as it relies to closed-form formulas. RDC starts with an empty projection matrix and in each iteration it adds one column with the goal of improving the verification performance. It converges when adding more columns does not yield further improvements. That is, its output is the most discriminative feature space it can achieve at the lowest dimension possible. Compared with other methods it does not appear to produce significant improvements in verification performance. However, it reduces the feature dimensionality by more than half in both paradigms, while outperforming the baseline. The high time cost reported in Table IV for RDC is mostly due to the fact that it minimizes the cost by considering all possible combinations of similar and dissimilar pairs. The authors have proposed a version of the algorithm that relies on ensemble learning to reduce the time cost. This version of the code, though, is not publicly available. Overall, methods that rely on pairwise relationships produce significant improvements in the restricted paradigm. This indicates that the information provided by pairwise relationships can be sufficient for distance metric learning methods to capture the semantics of what is considered to be similar or dissimilar. In fact, they seem to outperform methods that rely on labels. However, this performance gap is reduced when we move to the unrestricted paradigm. Hence, we conclude that the pairwise information provided in the restricted scenario is not sufficient to estimate the label information. To ensure that sufficient information is available for training the algorithms, the subsequent experiments are conducted using the unrestricted paradigm.

Experiment 2: The objective of this experiment is to assess the impact of the feature dimensionality on the verification performance and time cost for training the algorithms under evaluation. Specifically, the feature dimensionality was increased to 200, 300, 400, and 500, respectively, while all of the other settings remained the same with the unrestricted configuration of experiment 1. In Fig. 5, the mean AUC values obtained for all the algorithms are depicted to demonstrate the effect of the feature dimensionality. Detailed results about the verification performance and time cost at feature lengths 300 and 500 are reported in Table V. The term δACC denotes the relative performance obtained when the input dimensionality was increased from 100 to 300 and 500, respectively. For example, δACC is defined as $\frac{ACC_2 - ACC_1}{ACC_1}$, where ACC_2 is the mean accuracy for a given method achieved using 300 (500) dimensional features, and ACC_1 is the mean accuracy achieved by the same method using 100 dimensional features. As demonstrated, the verification performance of the baseline appears to increase as the feature dimensionality increases. This is an indication that higher dimensional features contain more discriminative information. The ACC and AUC of algorithms that rely on labels appear to increase with the feature dimensionality. The most significant improvements are obtained for GB-LMNN. Its ACC ranks second for feature dimentionality of 500. However, this is not the case for algorithms that rely on pairwise relationships. In particular, the verification performance of KISSME appears to decrease significantly when the feature dimensionality increases. The ACC and AUC performance for Sub-SML and DML-eig increases when the feature length is set to 300. However, the ACC and AUC performance decreases when the feature length is further increased to 500. Even though the verification performance of Sub-SML appears to decrease when the feature length is more than 200, it still achieves the best verification performance. Increasing the dimensionality of the input feature may have several impacts on the task of distance metric learning: (i) the signal to noise ratio of the PCA processed features might decrease; (ii) the numbers of parameters learned increases quadratically; and (iii) the learned metrics may over-fit the training data. In our experience, distance metric learning methods that rely on pairwise relationships appear to be better suited for face verification. The reason is that the corresponding evaluation protocol relies on one-to-one (i.e., pairwise) comparisons. The experimental results indicate that algorithms that rely on labels seem to be able to utilize the richer information provided by high dimensional features. However, they do not seem to be able to outperform the performance of methods developed using pairwise relationships (i.e., Sub-SML). On the other hand, algorithms that rely on pairwise relationships cannot reap the richer information of higher dimensional spaces. Specifically, increasing the feature length results in significantly degraded performance (i.e., KISSME). Another observation is that the average training time increases at a fast rate with the feature dimensionality. This is not the case for R-MLR and RDC. The training time for RDC depends partially on the output feature dimensionality, which does not increase linearly with the input feature length.

Experiment 3: In this experiment, we focus on distance metric learning algorithms that rely on pairwise relationships. Specifically, our goal is to assess the impact that the number of training pairs has on the verification performance and the training time cost. The dimensionality of the features was set to 100. The number of similar pairs generated was set to 600 and 700 per fold. In each case, the same number of dissimilar pairs was generated, respectively. The obtained results are summarized in Table VI. To analyze the impact of the number of pairwise relationships used for training we report the relative improvements obtained compared to experiment 1. For example, σACC is defined as $\frac{ACC_3 - ACC_1}{ACC_1}$, where ACC_3 is the mean accuracy achieved by a given method under the settings of experiment 3 (i.e., 600 or 700 pairs), and ACC_1 is the mean accuracy achieved by the same method under the settings of experiment 1 (i.e., 500 pairs). In general, the impact of the number of training pairs on the verification performance results in minor changes. For the baseline, using more training pairs yields a slightly worse estimation of the threshold. Consequently, the testing accuracy appears to drop. Since the testing set is not changed, the AUC remains the same. The ACC and AUC for DML-eig appear to benefit the most. This consistency is observed only for AUC when it comes to the other algorithms. That is, the class separation of the projected data appears to increase when the number of training pairs is increased for all methods. The results for ACC and VR are influenced by the estimation of the threshold. In summary, by increasing the number of training pairs a distance metric with better discriminative properties may be obtained. However, the experimental results do not provide sufficient evidence to draw definite conclusions.

Experiment 4: The objective of this experiment is to assess the generalization performance of the algorithms under evaluation. To this end, we report the training and testing verification performance under different settings (i.e., different feature dimensionality and different number of pairwise relationships used for training). Specifically, the *base configuration* is defined as the unrestricted paradigm of experiment 1. That is, the input dimensionality is set to 100 and the number of similar/dissimilar pairs is set to 500 for each fold. The results obtained for the base configuration are used as baseline to assess the relative impact of the feature dimensionality and the number of pairwise relationships. We define as *feature length configuration* the setting where the dimensionality of the features is increased to 500, while keeping the number of similar/dissimilar pairs at 500. Finally, we define as *pairs size configuration* the setting where the number of similar/dissimilar pairs is increased to 700 for each fold, while the dimensionality of features is kept at 100. A summary of the results obtained is offered in Table VII. The top part of the table reports the verification performance of the training phase in absolute values. Specifically, the subscript T is used to denote the results that correspond to the training phase. For example, ACC_T denotes the mean accuracy obtained for the training set. The bottom part of the table reports the verification performance of the testing phase in relative values. In particular, the prefix r is used to denote that the relative values correspond to the testing phase. For example, $rACC$ is defined as $\frac{ACC - ACC_T}{ACC_T}$, where ACC is the accuracy obtained in testing, and ACC_T is the accuracy obtained in training. The testing verification performance for the baseline appears to be better than the training one. However, when distance metric learning methods are employed the testing verification performance is degraded in all cases. For the base configuration, NCMML appears to generalize well as the gap between training and testing is the smallest. Nevertheless, its verification performance is the lowest in terms of absolute values (i.e., 80.61% accuracy in the training phase). When comparing the base and feature length configurations, we observe that the training verification for all the algorithms is increased. However, with the exception of GB-LMNN, the gap of the verification performance between the training and testing phases increases as well. For instance, the AAC_T for KISSME is close to 100% but the corresponding testing AUC is 85%. This provides us with strong evidence that KISSME over-fits the training data and explains why the verification performance observed in experiment 2 is degraded. We also observe that the performance gap seems to be smaller for algorithms that rely on labels. This seems to be one of the reasons why they can effectively utilize the richer discriminative information provided by high-dimensional input features. When moving from the base configuration to the pairs size configuration, we observe that the training verification performance drops but the gap between the training and testing performance is reduced. That is, when the number of training samples is increased, the corresponding algorithms fail to classify correctly each and every pair of the training samples. However, using an enhanced training set results in models that capture the statistical properties of the data in a better way, yielding better generalization performance. That is, the distance metric learning algorithms appear to capture the semantic context of similar/dissimilar human faces without over-fitting the data.

V. Conclusion

In this paper, we offered an overview of selected papers published during the years 2011-2013. Moreover, we proposed

TABLE VII

SUMMARY OF RESULTS FOR EXPERIMENT 4. THE TOP PART OF THE TABLE CONTAINS THE ABSOLUTE VALUES OBTAINED FOR THE TRAINING PHASE (ACC_T, VR_T, AND AUC_T), WHILE THE BOTTOM PART CONTAINS THE RELATIVE CHANGES OBTAINED FOR THE TESTING PHASE ($rACC$, rVR, AND $rAUC$). DIM. DENOTES THE INPUT FEATURE LENGTH AND NSPF DENOTES THE NUMBER OF SIMILAR/DISSIMILAR PAIRS PER FOLD.

Method	Base Configuration dim.: 100; NSPF: 500			Feature Length Configuration dim.: 300; NSPF: 500			Pairs Size Configuration dim.: 100; NSPF: 700		
	ACC_T (%)	VR_T (%)	AUC_T (%)	ACC_T (%)	VR_T (%)	AUC_T (%)	ACC_T (%)	VR_T (%)	AUC_T (%)
Baseline	69.32	63.46	76.12	70.96	64.65	77.93	69.25	63.39	76.16
DML-eig	84.58	82.86	92.48	95.61	94.16	99.06	83.84	82.58	91.93
KISSME	88.62	87.74	95.56	96.79	96.83	99.55	87.85	87.46	95.01
Sub-SML	**94.14**	**94.02**	**98.52**	**99.60**	**99.67**	**99.99**	**93.75**	**93.72**	**98.33**
RDC	83.06	84.17	91.53	91.51	91.93	97.39	81.41	80.98	90.03
R-MLR	89.37	77.21	86.13	93.61	77.76	86.42	-	-	-
NCMML	80.61	78.86	87.50	81.84	79.66	88.25	-	-	
BoostMetric	82.87	82.80	89.95	85.51	83.93	92.60	-	-	-
GB-LMNN	85.46	79.38	90.80	88.55	82.31	92.89	-	-	-
Method	$rACC$ (%)	rVR (%)	$rAUC$ (%)	$rACC$ (%)	rVR (%)	$rAUC$ (%)	$rACC$ (%)	rVR (%)	$rAUC$ (%)
Baseline	1.44	2.54	1.68	0.68	2.19	1.67	0.89	2.80	0.89
DML-eig	-8.43	-9.40	-6.82	-17.16	-21.30	-10.76	-6.61	-7.81	-6.61
KISSME	-8.63	-13.84	-6.00	**-25.28**	**-49.81**	**-14.62**	-6.94	-11.12	-6.94
Sub-SML	**-11.23**	-15.55	-6.29	-16.06	-25.62	-7.01	**-14.23**	**-22.66**	-6.43
RDC	-10.57	-9.98	**-10.10**	-18.35	-17.73	-14.56	-8.94	-8.78	**-8.94**
R-MLR	-8.88	-13.93	-6.16	-12.07	-20.91	-6.94	-	-	-
NCMML	-2.34	-4.55	-1.89	-2.67	-6.75	-1.87	-	-	-
BoostMetric	-5.19	-8.90	-4.28	-5.83	-12.75	-3.83	-	-	-
GB-LMNN	-8.97	**-17.70**	-6.77	-6.64	-12.43	-4.53	-	-	-

a taxonomy that organizes distance metric learning methods into five categories: (i) ensemble, (ii) non-linear, (iii) regularized, (iv) probabilistic, and (v) cost-variant. This way, we shed light on recently proposed approaches. More importantly, the proposed taxonomy can be used as a guide for future papers. That is, researchers can use the proposed categories to place their work in the right context and compare to the appropriate methods. Moreover, an empirical evaluation was conducted to analyze the performance of selected methods under the standard LFW protocols. Different settings were used to access: (i) the effect of the constraints used (i.e., pairwise vs. labels); (ii) the impact of the feature length and the number of training samples; and (iii) the generalization capabilities for the methods under evaluation. When sufficient information is provided for training, methods that rely on pairwise relationships yield comparable verification performance with methods that rely on labels. However, methods in the first category seem to be more efficient. Most of the distance metric learning algorithms fit the training data in a better way when the feature dimensionality is increased. Nevertheless, they fail to generalize equally well to the test set, which prevents them from fully utilizing the information contained in the high-dimensional features. This problem appears to be addressed by increasing the number of training samples. However, the number of similar pairwise relationships that LFW can yield is relatively small. Hence, the impact of the number of constraints on the verification performance is not as pronounced as the impact of the feature dimensionality. Distance metric learning methods could be improved in the future by utilizing the information of high-dimensional features more effectively.

ACKNOWLEDGMENT

The authors would like to thank Mrs. Siyang Wang for her contribution in finding relevant papers for this survey.

REFERENCES

[1] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[2] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07–49, Oct. 2007.

[3] L. Yang, "Distance metric learning: A comprehensive survey," Michigan State University, Tech. Rep., 2006.

[4] ——, "An overview of distance metric learning," in *Proc. IEEE Conference on Computer Vision and Pattern recognition*, Miami, FL, June 18–23 2007.

[5] D. Ramanan and S. Baker, "Local distance functions: A taxonomy, new algorithms, and an evaluation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 4, pp. 794–806, April 2011.

[6] K. Weinberger and L. Saul, "Distance metric learning for large margin nearest neighbor classification," *The Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.

[7] B. Kulis, "Metric learning: A survey," *Foundations & Trends in Machine Learning*, vol. 5, no. 4, pp. 287–364, 2012.

[8] A. Bellet, A. Habrard, and M. Sebban, "A survey on metric learning for feature vectors and structured data," University of Southern California, Los Angeles, CA 90089, Tech. Rep. arXiv:1306.6709v4, Aug. 19 2013.

[9] C.-C. Chang, "A boosting approach for supervised mahalanobis distance metric learning," *Pattern Recognition*, vol. 45, no. 2, pp. 844–862, 2012.

[10] C. Shen, J. Kim, L. Wang, and H. van den Hengel, "Positive semidefinite metric learning using boosting-like algorithms," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 1007–1036, 2012.

[11] J. Bi, D. Wu, L. Lu, M. Liu, Y. Tao, and M. Wolf, "Adaboost on low-rank psd matrices for metric learning," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 21–23 2011, pp. 2617–2624.

[12] T. Kozakaya, S. Ito, and S. Kubota, "Random ensemble metrics for object recognition," in *Proc. 13th IEEE International Conference on Computer Vision*, Barcelona, Spain, Nov. 6–13 2011, pp. 1959–1966.

[13] D. Kedem, S. Tyree, K. Weinberger, F. Sha, and G. Lanckriet, "Non-linear metric learning," in *Proc. 25th Annual Conference on Neural Information Processing Systems*, Lake Tahoe, NV, Dec. 3-6 2012, pp. 2582–2590.

[14] M. Norouzi, D. Fleet, and R. Salakhutdinov, "Hamming distance metric learning." in *Proc. 25th Annual Conference on Neural Information Processing Systems*, Lake Tahoe, NV, Dec. 5–8 2012, pp. 1070–1078.

[15] P. Jain, B. Kulis, J. Davis, and I. Dhillon, "Metric and kernel learning using a linear transformation," *The Journal of Machine Learning Research*, vol. 13, no. 1, pp. 519–547, 2012.

[16] A. Bellet, A. Habrard, and M. Sebban, "Similarity learning for provably accurate sparse linear classification," in *Proc. 29th International Conference on Machine Learning*, Edinburgh, Scotland, June 26 - July 1 2012.

[17] M. Liu and B. Vemuri, "A robust and efficient doubly regularized metric learning approach," in *Proc. 12th European Conference on Computer Vision*, Firenze, Italy, Oct. 7–13 2012, pp. 646–659.

[18] D. Lim, G. Lanckriet, and B. McFee, "Robust structural metric learning," in *Proc. 30th International Conference on Machine Learning*, Atlanta, GA, June 16–21 2013, pp. 615–623.

[19] N. Jiang, W. Liu, and Y. Wu, "Order determination and sparsity-regularized metric learning adaptive visual tracking," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, June 16–21 2012, pp. 1956–1963.

[20] Q. Cao, Y. Ying, and P. Li, "Similarity metric learning for face recognition," in *Proc. 14th IEEE International Conference on Computer Vision*, Sydney, Australia, Dec. 3–6 2013.

[21] M. Koestinger, M. Hirzer, P. Wohlhart, P. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, June 16–21, Providence, RI 2012.

[22] T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka, "Distance-based image classification: generalizing to new classes at near zero cost," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 11, pp. 2624–2637, Nov. 2013.

[23] A. Mignon and F. Jurie, "PCCA: A new approach for distance learning from sparse pairwise constraints," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, June 16–21 2012, pp. 2666–2672.

[24] W.-S. Zheng, S. Gong, and T. Xiang, "Reidentification by relative distance comparison," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 653–668, 2013.

[25] Y. Ying and P. Li, "Distance metric learning with eigenvalue optimization," *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 1–26, Jan. 13 2012.

[26] M. Kostinger, P. Wohlhart, P. Roth, and H. Bischof, "Joint learning of discriminative prototypes and large margin nearest neighbor classifiers," in *Proc. 14th IEEE International Conference on Computer Vision*, Sydney, Australia, Dec. 3–6 2013, pp. 3112–3119.

[27] S. Li and S. Shan, "Margin emphasized metric learning and its application to gabor feature based face recognition," in *Proc. IEEE 9th International Conference on Automatic Face and Gesture Recognition and Workshops*, Shanghai, China, March 22–24 2011, pp. 579–584.

[28] Y. Yu, J. Jiang, and L. Zhang, "Distance metric learning by minimal distance maximization," *Pattern Recognition*, vol. 44, no. 3, pp. 639–649, 2011.

[29] J. Lu, J. Hu, X. Zhou, Y. Shang, Y.-P. Tan, and G. Wang, "Neighborhood repulsed metric learning for kinship verification," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, June 16–21 2012, pp. 2594–2601.

[30] P. C. Mahalanobis, "On the generalized distance in statistics," *Proc. National Institute of Sciences*, vol. 2, pp. 49–55, 1936.

[31] M. Law, N. Thome, and M. Cord, "Quadruplet-wise image similarity learning," in *Proc. 14th IEEE International Conference on Computer Vision*, Sydney, Australia, Dec. 3–6 2013.

[32] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighbourhood components analysis," in *Proc. 18th Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, Dec. 13–18 2004.

[33] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, "Large scale online learning of image similarity through ranking," *The Journal of Machine Learning Research*, vol. 11, pp. 1109–1135, 2010.

[34] R. Schapire, "The strength of weak learnability," *Machine learning*, vol. 5, no. 2, pp. 197–227, 1990.

[35] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, pp. 123–140, 1996.

[36] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139, Aug. 1997.

[37] M.-F. Balcan, A. Blum, and N. Srebro, "Improved guarantees for learning via similarity functions," in *Proc. 21st Annual Conference on Learning Theory*, Helsinki, Finland, July 9–12 2008, pp. 287–298.

[38] R. Chatpatanasiri, T. Korsrilabutr, P. Tangchanachaianan, and B. Kijsirikul, "A new kernelization framework for mahalanobis distance learning algorithms," *Neurocomputing*, vol. 73, no. 10, pp. 1570–1579, 2010.

[39] B. McFee and G. Lanckriet, in *Proc. 27th International Conference on Machine Learning*, 2010, pp. 775–782.

[40] P. Belhumeur, J. Hespanha, and D. Kriegman, "Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711–720, 1997.

[41] A. R. Webb, *Statistical pattern recognition*. John Wiley & Sons, 2003.

[42] A. Izenman, *Linear Discriminant Analysis*. Springer, 2008.

[43] N. Jiang, W. Liu, and Y. Wu, "Adaptive and discriminative metric differential tracking," in *Proc. 24th IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 21–23 2011, pp. 1161–1168.

[44] Z. Hong, X. Mei, and D. Tao, "Dual-force metric learning for robust distracter-resistant tracker," in *Proc. 12th European Conference on Computer Vision*, Firenze, Italy, Oct. 7–13 2012, pp. 513–527.

[45] P. Yang, K. Huang, and C.-L. Liu, "Geometry preserving multi-task metric learning," *Machine learning*, vol. 92, no. 1, pp. 133–175, 2013.

[46] X. Li, C. Shen, Q. Shi, A. Dick, and A. van den Hengel, "Non-sparse linear representations for visual tracking with online reservoir metric learning," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, RI, June 16–21 2012, pp. 1760–1767.

[47] G. Niu, B. Dai, M. Yamada, and M. Sugiyama, "Information-theoretic semi-supervised metric learning via entropy regularization," in *Proc. 29th International Conference on Machine Learning*, Edinburgh, Scotland, June 26-July 1 2012.

[48] Q. Wang, P. Yuen, and G. Feng, "Semi-supervised metric learning via topology preserving multiple semi-supervised assumptions," *Pattern Recognition*, vol. 46, no. 9, pp. 2576–2587, 2013.

[49] F. Wang, "Semi-supervised metric learning by maximizing constraint margin," *IEEE Transactions on Cybernetics*, vol. 41, no. 4, pp. 931–939, 2011.

[50] B. Liu, M. Wang, R. Hong, Z. Zha, and X.-S. Hua, "Joint learning of labels and distance metric," *IEEE Transactions on Cybernetics*, vol. 40, no. 3, pp. 973–978, 2010.

[51] R. Cinbis, J. Verbeek, and C. Schmid, "Unsupervised metric learning for face identification in tv video," in *Proc. 13th IEEE International Conference on Computer Vision*, Barcelona, Spain, Nov. 6–13 2011, pp. 1559–1566.

[52] X. Cao, D. Wipf, F. Wen, G. Duan, and J. Sun, "A practical transfer learning algorithm for face verification," in *Proc. 14th IEEE International Conference on Computer Vision*, Sydney, Australia, Dec. 3–6 2013.

[53] S. Wang, S. Jiang, Q. Huang, and Q. Tian, "Multi-feature metric learning with knowledge transfer among semantics and social tagging," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, RI, June 16–21 2012, pp. 2240–2247.

[54] D. Ryabko and J. Mary, "A binary-classification-based metric between time-series distributions and its use in statistical and learning problems," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 2837–2856, 2013.

[55] A. Bellet, A. Habrard, and M. Sebban, "Good edit similarity learning by loss minimization," *Machine Learning*, pp. 1–31, 2012.

[56] B. Shaw, B. Huang, and T. Jebara, "Learning a distance metric from a network." in *Proc. 25th Annual Conference on Neural Information Processing Systems*, Granada, Spain, Dec. 12-15 2011, pp. 1899–1907.

[57] A. Slivkins, "Multi-armed bandits on implicit metric spaces." in *Proc. 25th Annual Conference on Neural Information Processing Systems*, Granada, Spain, Dec. 1215 2011, pp. 1602–1610.

[58] F. Schroff, T. Treibitz, D. Kriegman, and S. Belongie, "Pose, illumination and expression invariant pairwise face-similarity measure via doppelgänger list comparison," in *Proc. 13th IEEE International Conference on Computer Vision*, Barcelona, Spain, Nov. 6–13 2011, pp. 2494–2501.

[59] J. Yi, R. Jin, A. Jain, S. Jain, and T. Yang, "Semi-crowdsourced clustering: generalizing crowd labeling by robust distance metric learning."

in *Proc. 26th Annual Conference on Neural Information Processing Systems*, Montreal, Canada, Dec. 8–11 2012, pp. 1781–1789.

[60] Y. Verma and C. Jawahar, "Image annotation using metric learning in semantic neighbourhoods," in *Proc. 12th European Conference on Computer Vision*, Firenze, Italy, Oct. 7–13 2012, pp. 836–849.

[61] C. Ji, X. Zhou, L. Lin, and W. Yang, "Labeling images by integrating sparse multiple distance learning and semantic context modeling," in *Proc. 12th European Conference on Computer Vision*, Firenze, Italy, Oct. 7–13 2012, pp. 688–701.

[62] Z. Feng, R. Jin, and A. Jain, "Large-scale image annotation by efficient and robust kernel metric learning," in *Proc.* 13^{th} *IEEE International Conference on Computer Vision*, Barcelona, Spain, Nov. 6–13 2011.

[63] N. Verma, D. Mahajan, S. Sellamanickam, and V. Nair, "Learning hierarchical similarity metrics," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, June 16–21 2012, pp. 2280–2287.

[64] N. Quadrianto and C. Lampert, "Learning multi-view neighborhood preserving projections," in *Proc. 28th International Conference on Machine Learning*, Bellevue, WA, June 28 – July 2 2011, pp. 425–432.

[65] H. Wang, F. Nie, and H. Huang, "Robust and discriminative distance for multi-instance learning," in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Providence, RI, June 16–21 2012.

[66] B. McFee and G. Lanckriet, "Learning multi-modal similarity," *The Journal of Machine Learning Research*, vol. 12, pp. 491–523, 2011.

[67] L. Cheng, "Riemannian similarity learning," in *Proc. 30th International Conference on Machine Learning*, Atlanta, GE, June 16–21 2013, pp. 540–548.

[68] S. Hauberg, O. Freifeld, and M. Black, "A geometric take on metric learning." in *Proc. 26th Annual Conference on Neural Information Processing Systems*, Montreal, Canada, Dec. 8–11 2012, pp. 2033–2041.

[69] N. Fan, "Learning nonlinear distance functions using neural network for regression with application to robust human age estimation," in *Proc. 13th IEEE International Conference on Computer Vision*, Barcelona, Spain, Nov. 6–13 2011, pp. 249–254.

[70] P. Zhu, L. Zhang, W. Zuo, and D. Zhang, "From point to set: extend the learning of distance metrics," in *Proc. 14th IEEE International Conference on Computer Vision*, vol. 50, no. 100, Sydney, Australia, Dec. 3–6 2013, p. 200.

[71] J. Lu, G. Wang, and P. Moulin, "Image set classification using holistic multiple order statistics features and localized multi-kernel metric learning," in *Proc. 14th IEEE International Conference on Computer Vision*, Sydney, Australia, Dec. 3–6 2013.

[72] Z. Wang, Y. Hu, and L.-T. Chia, "Improved learning of i2c distance and accelerating the neighborhood search for image classification," *Pattern Recognition*, vol. 44, no. 10, pp. 2384–2394, 2011.

[73] Z. Wang, S. Gao, and L.-T. Chia, "Learning class-to-image distance via large margin and l1-norm regularization," in *Proc. 12th European Conference on Computer Vision*, Firenze, Italy, Oct. 7–13 2012, pp. 230–244.

[74] R. Weng, J. Lu, J. Hu, G. Yang, and Y.-P. Tan, "Robust feature set matching for partial face recognition," in *Proc. 14th IEEE International Conference on Computer Vision*, Sydney, Australia, Dec. 3–6 2013.

[75] Y. Wu, M. Minoh, M. Mukunoki, and S. Lao, "Set based discriminative ranking for recognition," in *Proc. 12th European Conference on Computer Vision*, Firenze, Italy, Oct. 7–13 2012, pp. 497–510.

[76] N. Verma, "Distance preserving embeddings for general n-dimensional manifolds," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 2415–2448, 2013.

[77] A. Cherian, S. Sra, A. Banerjee, and N. Papanikolopoulos, "Efficient similarity search for covariance matrices via the jensen-bregman logdet divergence," in *Proc. 13th IEEE International Conference on Computer Vision*, Barcelona, Spain, Nov. 6–13 2011, pp. 2399–2406.

[78] C. Fang and L. Torresani, "Measuring image distances via embedding in a semantic manifold," in *Proc. 12th European Conference on Computer Vision*, Firenze, Italy, Oct. 7-13 2012, pp. 402–415.

[79] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun, "Bayesian face revisited: A joint formulation," in *Proc. 12th European Conference on Computer Vision*, Firenze, Italy, Oct. 7–13 2012, pp. 566–579.

[80] M. Der and L. Saul, "Latent coincidence analysis: A hidden variable model for distance metric learning," in *Proc. Advances in Neural Information Processing Systems*, Lake Tahoe, NV, Dec. 3–8 2012, pp. 3239–3247.

[81] S. Changpinyo, K. Liu, and F. Sha, "Similarity component analysis," in *Proc. 27th Annual Conference on Neural Information Processing Systems*, Lake Tahoe, NV, Dec. 5–8 2013, pp. 1511–1519.

[82] M. Titsias and M. Lázaro-Gredilla, "Variational inference for mahalanobis distance metrics in gaussian process regression," in *Proc. 27th Annual Conference on Neural Information Processing Systems*, Lake Tahoe, NV, Dec. 5–8 2013, pp. 279–287.

[83] Y. Hwang and H.-K. Ahn, "Convergent bounds on the euclidean distance." in *Proc. 25th Annual Conference on Neural Information Processing Systems*, Granada, Spain, Dec. 12-15 2011, pp. 388–396.

[84] P. Kar and P. Jain, "Similarity-based learning via data driven embeddings." in *Proc. 25th Annual Conference on Neural Information Processing Systems*, Granada, Spain, Dec. 1215 2011, pp. 1998–2006.

[85] J. Yang, L. Zhang, J.-Y. Yang, and D. Zhang, "From classifiers to discriminators: a nearest neighbor rule induced discriminant analysis," *Pattern recognition*, vol. 44, no. 7, pp. 1387–1402, 2011.

[86] Y. Hong, Q. Li, J. Jiang, and Z. Tu, "Learning a mixture of sparse distance metrics for classification and dimensionality reduction," in *Proc. IEEE 13th International Conference on Computer Vision*, Barcelona, Spain, Nov. 13–16 2011, pp. 906–913.

[87] J. Wang, H. Do, A. Woznica, and A. Kalousis, "Metric learning with multiple kernels." in *Proc. 25th Annual Conference on Neural Information Processing Systems*, Granada, Spain, Dec. 1215 2011, pp. 1170–1178.

[88] J. Wang, A. Woznica, and A. Kalousis, "Parametric local metric learning for nearest neighbor classification," in *Proc. 25th Annual Conference on Neural Information Processing Systems*, Lake Tahoe, NV, Dec. 3–6 2012, pp. 1–9.

[89] C. Shen, J. Kim, and L. Wang, "A scalable dual approach to semidefinite metric learning," in *Proc. 24th IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 21–23 2011, pp. 2601–2608.

[90] K. Park, C. Shen, Z. Hao, and J. Kim, "Efficiently learning a distance metric for large margin nearest neighbor classification," in *Proc. 25th Conference on Artificial Intelligence*, San Francisco, CA, Aug. 711 2011.

[91] M. Guillaumin, J. Verbeek, and C. Schmid, "Is that you? metric learning approaches for face identification," in *Proc. 12th IEEE International Conference on Computer Vision*, Kyoto, Japan, Sept. 27–Oct. 4 2009.

[92] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the Wild: A database for studying face recognition in unconstrained environments," in *Proc. European Conference on Computer Vision Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, Marseille, France, Oct. 17–20 2008.

[93] P. Li, Y. Fu, U. Mohammed, J. Elder, and S. Prince, "Probabilistic models for inference about identity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 144–157, 2012.

[94] M. Everingham, J. Sivic, and A. Zisserman, "Hello! my name is... buffy-automatic naming of characters in tv video," in *Proc. 17th British Machine Vision Conference*, Edinburgh, Scotland, Sept. 4–7 2006.

[95] I. Jolliffe, *Principal component analysis*. Springer Verlag, New York, 1986.