mathematics

MDPI

Article

Multi-View Cosine Similarity Learning with Application to Face Verification

Zining Wang [1], Jiawei Chen [2] and Junlin Hu [1,*]

[1] School of Software, Beihang University, Beijing 100191, China; 19373122@buaa.edu.cn
[2] College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China; 2019210507@buct.edu.cn
* Correspondence: hujunlin@buaa.edu.cn

Abstract: An instance can be easily depicted from different views in pattern recognition, and it is desirable to exploit the information of these views to complement each other. However, most of the metric learning or similarity learning methods are developed for single-view feature representation over the past two decades, which is not suitable for dealing with multi-view data directly. In this paper, we propose a multi-view cosine similarity learning (MVCSL) approach to efficiently utilize multi-view data and apply it for face verification. The proposed MVCSL method is able to leverage both the common information of multi-view data and the private information of each view, which jointly learns a cosine similarity for each view in the transformed subspace and integrates the cosine similarities of all the views in a unified framework. Specifically, MVCSL employs the constraints that the joint cosine similarity of positive pairs is greater than that of negative pairs. Experiments on fine-grained face verification and kinship verification tasks demonstrate the superiority of our MVCSL approach.

Keywords: similarity learning; metric learning; multi-view; cosine similarity; face verification

MSC: 62H20; 62H30; 68T10

Citation: Wang, Z.; Chen, J.; Hu, J. Multi-View Cosine Similarity Learning with Application to Face Verification. *Mathematics* **2022**, *10*, 1800. https://doi.org/10.3390/math10111800

Academic Editor: Radu Tudor Ionescu

Received: 24 April 2022
Accepted: 23 May 2022
Published: 25 May 2022

1. Introduction

Metric learning or similarity learning aims to develop an effective metric to measure the similarities of samples [1]. Samples from the same class are projected into neighboring locations in the embedding space while samples of various categories are separated. Recently, numerous metric learning or similarity learning approaches have been introduced [1–3] and they have achieved a great success for numerous visual understanding tasks including face verification [2], image retrieval [3], image classification [4], and person re-identification.

Face verification is a representative task of pattern recognition and computer vision; its purpose is to decide whether a pair of facial images belongs to the same subject or not. Face verification has received wide attention since the unconstrained face image datasets were released to the public, for example, labeled faces in the wild (LFW) (LFW) [5], MegaFace [6] and other benchmark face image datasets [7]. A variety of metric learning-based face verification methods have been introduced in the literature [2,8] to advance the performance of face verification. Guillaumin et al. [9] proposed a logistic discriminant method and a nearest neighbor method to learn a distance metric for calculating the similarity of two face images. Koestinger et al. [10] introduced a large-scale metric learning method to compute the Mahalanobis distance of images from the statistical inference perspective and achieved the state-of-the-art performance. Schroff et al. [2] exploited the deep convolutional neural networks for face verification and clustering and proposed a FaceNet method to measure the distance of face images in Euclidean space. More detailed introductions and developments of face verification can refer to the survey paper [8].

Most of the metric learning approaches in face verification are developed for single-view data so that they are not suitable for exploiting multi-view data efficiently. Multi-view data are very common in the practical applications, and it usually describes the information of the examples more comprehensively than single-view data. For instance, we can use different feature representations to depict a face image, e.g., scale invariant feature transform (SIFT) [11], local binary pattern (LBP) [12] and histogram of oriented gradient (HOG) [13]. Multi-view learning aims to improve the performance of the classification or recognition tasks by making use of multi-view representations of data. For the sake of utilizing multi-view data, many multi-view learning methods have been introduced in the last decade [14,15]; however, there are only a small number of them developed in the multi-view metric learning perspective, and the existing multi-view metric learning methods are mainly formulated in the framework of Mahalanbis distance metric learning.

In this paper, we develop a multi-view cosine similarity learning (MVCSL) approach to efficiently utilize multi-view data from the cosine similarity learning framework. To capture the correlation across multiple views and exploit the private information of each view, the proposed MVCSL method jointly learns a cosine similarity for each view in the transformed subspace and seeks the optimal combination of multiple cosine similarities of multi-view feature representations under a unified framework, where the joint cosine similarity of each positive pair is forced to be greater than a large constant value and the joint cosine similarity of each negative pair is forced to be less than a small threshold. Experimental results on the fine-grained face verification and facial kinship verification applications demonstrate the advantages of our MVCSL method. Figure 1 illustrates the basic idea of our proposed MVCSL approach.

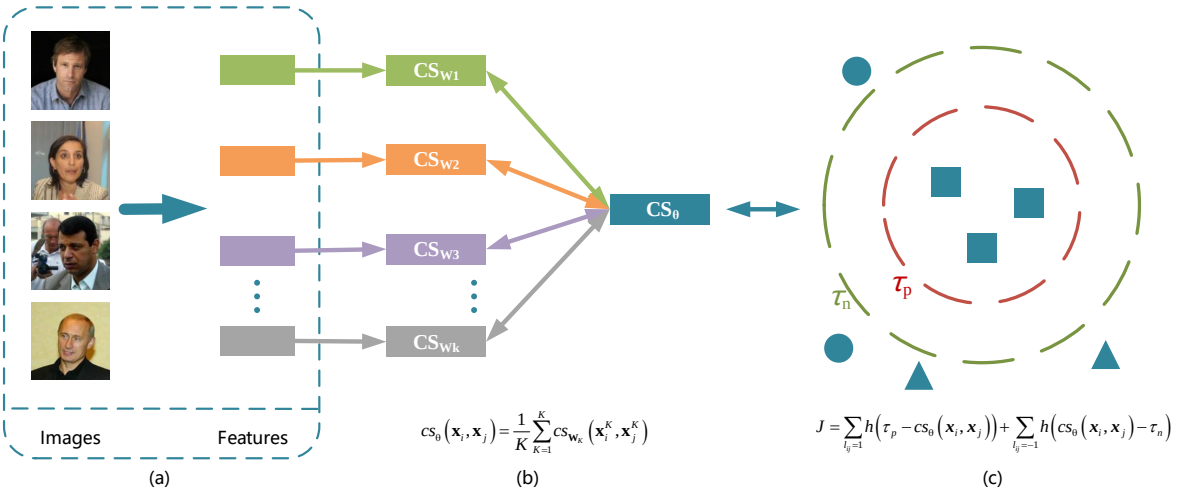

Figure 1. Illustration of the basic ideas of the proposed MVCSL method. (**a**) Preprocessing, (**b**) cosine metric, (**c**) MVCSL. Given multi-view feature representations of each sample, MVCSL under the large margin framework learns the optimal combination of the cosine similarities of multi-view data, which constrains the joint cosine similarity of positive samples to be greater than a large value t_p and that of negative samples to be less than a small value t_n.

The remainder of this paper is structured as follows. Section 2 briefly reviews the related work. In Section 3, we detail the proposed multi-view cosine similarity learning approach, and the experiments for face verification are presented in Section 4. Finally, the paper is concluded in Section 5.

2. Related Work

The target of metric learning or similarity learning is, broadly speaking, to learn a proper similarity measure for increasing the dissimilarity of inter-class samples and increasing the similarity of intra-class samples. A variety of metric learning approaches have been

proposed over the past decade, and they have been used in various applications of pattern recognition including face recognition, image searching and fine-grained recognition. For example, Xing et al. [16] designed metric learning as a convex optimization problem by adopting a semidefinite programming formulation of similarity side information. Weinberger et al. [17] introduced the classical large margin nearest neighbor (LMNN) algorithm that makes samples from the same class compose k-nearest neighbors and samples belonging to various categories be separated by an appropriate margin. Davis et al. [18] presented the information theoretic metric learning (ITML) approach by maximizing the entropy of a multivariate Gaussian to learn a Mahalanobis distance. The keep it simple and straightforward metric learning (KISSME) [10] method was proposed to learn a distance metric by the maximum likelihood estimators from the perspective of the statistical inference. Nguyen and Bai [19] introduced a cosine similarity metric learning (CSML) method using the cosine distance as the similarity measurement for a face verification task. In addition, several methods based on fractal theory [20–22] were introduced to find an appropriate distance metric for face recognition. Over the past few years, with the prosperity of deep learning algorithms, more and more deep metric learning approaches [2,3,23] were also presented to learn the nonlinear mapping functions using deep neural networks.

Although metric learning methods have been developed so far, most of them primarily aim to seek a metric or similarity function for either the single-view feature or cascading multiple types of features so that they cannot efficiently exploit multi-view feature representations. For the sake of better exploiting multi-view data that usually include the complementary information, several multi-view metric learning methods [7,15,24–26] have been introduced to learn a more comprehensive metric than the single-view based metric learning approaches. For instance, Lu et al. [7] developed a multi-view neighborhood repulsed metric learning approach to utilize multiple feature representations of samples for a kinship verification task. Xie and Xing [24] introduced a multi-modal distance metric learning method that maps the samples in a single latent feature space. Hu et al. [15] proposed a sharable and individual multi-view metric learning method to make use of both the private characteristics from each view and the shared representation for different views. Jia et al. [26] introduced a semi-supervised multi-view deep discriminant representation learning method, which utilizes the consensus content of inter-view features and reduces the redundancy of feature representations. However, these existing multi-view metric learning methods are mainly formulated in the framework of Mahalanbis distance metric learning. In this paper, we present a multi-view cosine similarity learning approach from the cosine similarity learning framework by collaboratively learning multiple cosine similarities to better exploit complementary information of multi-view feature representations.

3. Multi-View Cosine Similarity Learning

Suppose that we have a training set with N training samples $\mathcal{X} = \{\mathbf{x}_i \in \mathbb{R}^q | i = 1, 2, \cdots, N\}$, where q is the dimension of the sample $\mathbf{x}_i$. For any sample, it can be easily depicted in multiple views with various feature representations. Let $\mathcal{X}_\kappa = \{\mathbf{x}_i^\kappa \in \mathbb{R}^{q_\kappa} | 1 \leq \kappa \leq K\}_{i=1}^N$ denote the features set of $\mathcal{X}$ from the κ-th view, where $\mathbf{x}_i^\kappa$ is the κ-th view representation of $\mathbf{x}_i$, and K and q_κ are the total number of views and the dimension of $\mathbf{x}_i^\kappa$, respectively.

In general, it is not desirable to directly map multi-view features into a unified subspace, because the distributions of different views are different in their independent subspaces, and it cannot take advantage of the specific characteristics of each view and ignores the differences of information among the various views. To overcome this limitation, we map samples of each view into the individual space via the linear transformation $\mathbf{W}_\kappa$, and the cosine similarity between $\mathbf{x}_i^\kappa$ and $\mathbf{x}_j^\kappa$ in the κ-th view is computed by:

$$cs_{\mathbf{W}_\kappa}(\mathbf{x}_i^\kappa, \mathbf{x}_j^\kappa) = \frac{(\mathbf{x}_i^\kappa)^\top \mathbf{W}_\kappa \mathbf{W}_\kappa^\top \mathbf{x}_j^\kappa}{\sqrt{(\mathbf{x}_i^\kappa)^\top \mathbf{W}_\kappa \mathbf{W}_\kappa^\top \mathbf{x}_i^\kappa} \sqrt{(\mathbf{x}_j^\kappa)^\top \mathbf{W}_\kappa \mathbf{W}_\kappa^\top \mathbf{x}_j^\kappa}}. \quad (1)$$

Considering that different view representations of each sample depict the same subject and they are able to complement each other with the difference information, the joint cosine similarity between the samples $\mathbf{x}_i$ and $\mathbf{x}_j$ is written as:

$$cs_\theta(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{K}\sum_{\kappa=1}^{K} cs_{\mathbf{W}_\kappa}(\mathbf{x}_i^\kappa, \mathbf{x}_j^\kappa), \tag{2}$$

where $\theta = \{\mathbf{W}_\kappa\}_{\kappa=1}^K$ and $cs_{\mathbf{W}_\kappa}(\mathbf{x}_i^\kappa, \mathbf{x}_j^\kappa)$ are the cosine similarity of the κ-th view between $\mathbf{x}_i$ and $\mathbf{x}_j$. Obviously, the cosine similarity score varies from -1 to 1, so it is very suitable for similarity learning.

We formulate our multi-view cosine similarity learning (MVCSL) method under the large margin framework to learn the optimal parameter $\theta = \{\mathbf{W}_\kappa\}_{\kappa=1}^K$. The objective function of our proposed MVCSL method is as:

$$\begin{aligned} \min\ J = &\sum_{l_{ij}=1} h\left(\tau_p - \frac{1}{K}\sum_{\kappa=1}^{K} cs_{\mathbf{W}_\kappa}(\mathbf{x}_i^\kappa, \mathbf{x}_j^\kappa)\right) \\ &+ \sum_{l_{ij}=-1} h\left(\frac{1}{K}\sum_{\kappa=1}^{K} cs_{\mathbf{W}_\kappa}(\mathbf{x}_i^\kappa, \mathbf{x}_j^\kappa) - \tau_n\right) \\ &+ \eta\sum_{\kappa=1}^{K}\|\mathbf{W}_\kappa - \mathbf{W}_0\|^2, \end{aligned} \tag{3}$$

in which η is the coefficient of the regularization term, and $h(x) = \max(x, 0)$. τ_p and τ_n are thresholds of cosine similarity for positive samples and negative samples, respectively, $-1 \le \tau_n \le \tau_p \le 1$. $\mathbf{W}_0$ is a transformation matrix with ones on the diagonal and zeros elsewhere. We treat the derivative $h'(0) = 0$ at point $x = 0$. Pairwise label $l_{ij} = 1$ represents that $\mathbf{x}_i$ and $\mathbf{x}_j$ come from the same object (i.e., positive pairs) and $l_{ij} = -1$ denotes that $\mathbf{x}_i$ and $\mathbf{x}_j$ are from different objects (i.e., negative pairs). By setting appropriate thresholds, the joint cosine similarity of positive samples is more than a large value τ_p; simultaneously, the joint cosine similarity between negative samples is less than a small value τ_n.

The gradient of J with regard to $\mathbf{W}_\kappa$ is calculated by:

$$\begin{aligned} \frac{\partial J}{\partial \mathbf{W}_\kappa} = &-\frac{1}{K}\sum_{l_{ij}=1} h'\left(\tau_p - cs_\theta(\mathbf{x}_i, \mathbf{x}_j)\right)\frac{\partial cs_{\mathbf{W}_\kappa}(\mathbf{x}_i^\kappa, \mathbf{x}_j^\kappa)}{\partial \mathbf{W}_\kappa} \\ &+ \frac{1}{K}\sum_{l_{ij}=-1} h'\left(cs_\theta(\mathbf{x}_i, \mathbf{x}_j) - \tau_n\right)\frac{\partial cs_{\mathbf{W}_\kappa}(\mathbf{x}_i^\kappa, \mathbf{x}_j^\kappa)}{\partial \mathbf{W}_\kappa} \\ &+ 2\eta\sum_{\kappa=1}^{K}(\mathbf{W}_\kappa - \mathbf{W}_0), \end{aligned} \tag{4}$$

in which the gradients of the cosine similarity $cs_{\mathbf{W}_\kappa}$ with regard to $\mathbf{W}_\kappa$ are computed by:

$$\begin{aligned} \frac{\partial cs_{\mathbf{W}_\kappa}\left(\mathbf{x}_i^\kappa, \mathbf{x}_j^\kappa\right)}{\partial \mathbf{W}_\kappa} &= \frac{\partial\left(\frac{f(\mathbf{W}_\kappa)}{g(\mathbf{W}_\kappa)}\right)}{\partial \mathbf{W}_\kappa} \\ &= \frac{1}{g(\mathbf{W}_\kappa)}\frac{\partial f(\mathbf{W}_\kappa)}{\partial \mathbf{W}_\kappa} - \frac{f(\mathbf{W}_\kappa)}{g^2(\mathbf{W}_\kappa)}\frac{\partial g(\mathbf{W}_\kappa)}{\partial \mathbf{W}_\kappa}, \end{aligned} \tag{5}$$

in which

$$\frac{\partial f(\mathbf{W}_\kappa)}{\partial \mathbf{W}_\kappa} = \frac{\partial\left((\mathbf{x}_i^\kappa)^\top \mathbf{W}_\kappa \mathbf{W}_\kappa^\top \mathbf{x}_j^\kappa\right)}{\partial \mathbf{W}_\kappa} = \left(\mathbf{x}_i^\kappa(\mathbf{x}_j^\kappa)^\top + \mathbf{x}_j^\kappa(\mathbf{x}_i^\kappa)^\top\right)\mathbf{W}_\kappa, \tag{6}$$

$$\begin{aligned}\frac{\partial g(\mathbf{W}_\kappa)}{\partial \mathbf{W}_\kappa} =& \frac{\partial\left(\sqrt{\left((\mathbf{x}_i^\kappa)^\top \mathbf{W}_\kappa \mathbf{W}_\kappa^\top \mathbf{x}_i^\kappa\right)\left((\mathbf{x}_j^\kappa)^\top \mathbf{W}_\kappa \mathbf{W}_\kappa^\top \mathbf{x}_j^\kappa\right)}\right)}{\partial \mathbf{W}_\kappa} \\ =& \frac{\sqrt{(\mathbf{x}_j^\kappa)^\top \mathbf{W}_\kappa \mathbf{W}_\kappa^\top \mathbf{x}_j^\kappa}}{\sqrt{(\mathbf{x}_i^\kappa)^\top \mathbf{W}_\kappa \mathbf{W}_\kappa^\top \mathbf{x}_i^\kappa}} \mathbf{x}_i^\kappa (\mathbf{x}_i^\kappa)^\top \mathbf{W}_\kappa \\ &+ \frac{\sqrt{(\mathbf{x}_i^\kappa)^\top \mathbf{W}_\kappa \mathbf{W}_\kappa^\top \mathbf{x}_i^\kappa}}{\sqrt{(\mathbf{x}_j^\kappa)^\top \mathbf{W}_\kappa \mathbf{W}_\kappa^\top \mathbf{x}_j^\kappa}} \mathbf{x}_j^\kappa (\mathbf{x}_j^\kappa)^\top \mathbf{W}_\kappa.\end{aligned} \tag{7}$$

Substituting Formulas (6) and (7) into Formula (5), we can obtain the gradient:

$$\begin{aligned}&\frac{\partial cs_{\mathbf{W}_\kappa}\left(\mathbf{x}_i^\kappa, \mathbf{x}_j^\kappa\right)}{\partial \mathbf{W}_\kappa} \\ &= \frac{\left(\mathbf{x}_i^\kappa (\mathbf{x}_j^\kappa)^\top + \mathbf{x}_j^\kappa (\mathbf{x}_i^\kappa)^\top\right)\mathbf{W}_\kappa}{\sqrt{(\mathbf{x}_i^\kappa)^\top \mathbf{W}_\kappa \mathbf{W}_\kappa^\top \mathbf{x}_i^\kappa}\sqrt{(\mathbf{x}_j^\kappa)^\top \mathbf{W}_\kappa \mathbf{W}_\kappa^\top \mathbf{x}_j^\kappa}} \\ &- \frac{(\mathbf{x}_i^\kappa)^\top \mathbf{W}_\kappa \mathbf{W}_\kappa^\top \mathbf{x}_j^\kappa}{\left((\mathbf{x}_i^\kappa)^\top \mathbf{W}_\kappa \mathbf{W}_\kappa^\top \mathbf{x}_i^\kappa\right)^{\frac{3}{2}}\left((\mathbf{x}_j^\kappa)^\top \mathbf{W}_\kappa \mathbf{W}_\kappa^\top \mathbf{x}_j^\kappa\right)^{\frac{1}{2}}} \mathbf{x}_i^\kappa (\mathbf{x}_i^\kappa)^\top \mathbf{W}_\kappa \\ &- \frac{(\mathbf{x}_i^\kappa)^\top \mathbf{W}_\kappa \mathbf{W}_\kappa^\top \mathbf{x}_j^\kappa}{\left((\mathbf{x}_i^\kappa)^\top \mathbf{W}_\kappa \mathbf{W}_\kappa^\top \mathbf{x}_i^\kappa\right)^{\frac{1}{2}}\left((\mathbf{x}_j^\kappa)^\top \mathbf{W}_\kappa \mathbf{W}_\kappa^\top \mathbf{x}_j^\kappa\right)^{\frac{3}{2}}} \mathbf{x}_j^\kappa (\mathbf{x}_j^\kappa)^\top \mathbf{W}_\kappa.\end{aligned} \tag{8}$$

After obtaining the gradient $\frac{\partial J}{\partial \mathbf{W}_\kappa}$, the stochastic gradient descent method is used to update $\mathbf{W}_\kappa$ iteratively for each view until the objective function of our MVCSL method is converged:

$$\mathbf{W}_\kappa = \mathbf{W}_\kappa - \mu \frac{\partial J}{\partial \mathbf{W}_\kappa}, \tag{9}$$

where μ is the learning rate, $\kappa = 1, 2, \ldots, K$. Algorithm 1 summarizes the main steps of the MVCSL approach, in which we initialize the κ-th transformation $\mathbf{W}_\kappa$ as a matrix with ones on the diagonal and zeros elsewhere, $\kappa = 1, 2, \ldots, K$.

Algorithm 1: MVCSL

Input: Training set $\mathcal{X}_\kappa = \left\{\mathbf{x}_i^\kappa \in \mathbb{R}^{q_\kappa}\right\}_{i=1}^N$ of the κ-th view; thresholds τ_p and τ_n; learning rate μ; total iterative number T; convergence error ξ.
Output: $\{\mathbf{W}_\kappa\}_{\kappa=1}^K$.

Initialize $\{\mathbf{W}_\kappa\}_{\kappa=1}^K$
Compute the initial J_0 by (3)
for $t = 1, 2, \ldots, T$ **do**
  Compute the joint similarity $cs_\theta(\mathbf{x}_i, \mathbf{x}_j)$ by (2)
  for $\kappa = 1, 2, \ldots, K$ **do**
    $\mathbf{W}_\kappa \leftarrow \mathbf{W}_\kappa - \mu \frac{\partial J}{\partial \mathbf{W}_\kappa}$
  end for
  Update J_t by (3)
  if $|J_t - J_{t-1}| < \xi$ **then**
    return $\{\mathbf{W}_\kappa\}_{\kappa=1}^K$
  end if
end for
return $\{\mathbf{W}_\kappa\}_{\kappa=1}^K$

4. Experiments

This section conducts experiments on fine-grained face verification and kinship verification to demonstrate the advantages of our MVCSL for exploiting multi-view data.

Following the common settings [15], we evaluate the proposed methods with three different similarity learning baseline approaches as:

- MVC-s: This is the single-view cosine similarity learning method that learns a single similarity metric via the objective function (3) using the single-view feature representation;
- Concatenation (abbrev., Con): All the multi-view feature representations are concatenated as a high-dimension feature vector, and then, the MVC-s method is employed to find out the cosine similarity;
- MVC-i: We independently learn the mapping for each view, and then, we add up the cosine similarities of all views as the final cosine similarity of a sample pair.

For the parameter settings of our MVCSL and baseline methods, we empirically set thresholds t_p and t_n as 0.8 and 0.1, and μ as 0.01 for all experiments.

4.1. Fine-Grained Face Verification

Fine-grained face verification is a special task in face verification, where each negative sample pair consists of very similar face images such as face images from twins and similarly looking facial image of different subjects, so it is more difficult than general face verification in real-world scenarios.

4.1.1. Dataset and Settings

The fine-grained face verification (FGFV) [27] dataset consists of 1820 face images categorized into 455 negative pairs of face images and 455 positive pairs of face images. The 455 negative pairs are collected from very similar face images of twins without restrictions on disturbances including lighting, expression, pose background and so on. The 455 positive pairs of face images are chosen from positive pairs of the LFW dataset [5]. We evaluate our proposed MVCSL and baseline approaches on the well-aligned version of the FGFV dataset, in which facial images were aligned and cropped into the pixels of 64×64. We then convert all images into the gray-scale and extract three hand-crafted features for each face image as:

- LBP [12]: we partition an image into 8×8 segments and obtain a 59-dimensional LBP for each segment; then, we finally achieve a 3776-dimensional feature representation by concatenating them.
- HOG [13]: we split an image into non-overlapping blocks 4×4 and 8×8 with two different sizes and compute a nine-dimensional HOG feature on each block. Finally, we achieve a feature representation of 2880 dimensions for each image.
- SIFT [11]: each facial image is segmented into 49 blocks to extract a feature representation of 6272 dimensions.

Lastly, each feature representation is reduced 200 dimensions by the principal component analysis (PCA) method. We employ a 5-fold cross-validation strategy to evaluate our method on the FGFV dataset under the image restricted setting, which only exploits the pairwise labels of positive pairs and negative pairs.

The fine-grained LFW (FGLFW) [28] dataset includes 10-fold face image pairs, and every fold is composed of 300 positive pairs of face images and 300 negative face image pairs. The positive face pairs are the same as LFW [5], but the negative pairs are similar face images that were manually selected from the LFW dataset. Figure 2 presents several negative pairs of face images from LFW, FGFV and FGLFW datasets, where the negative pairs of FGFV and FGLFW datasets are easy to incorrectly identify as the positive pairs.

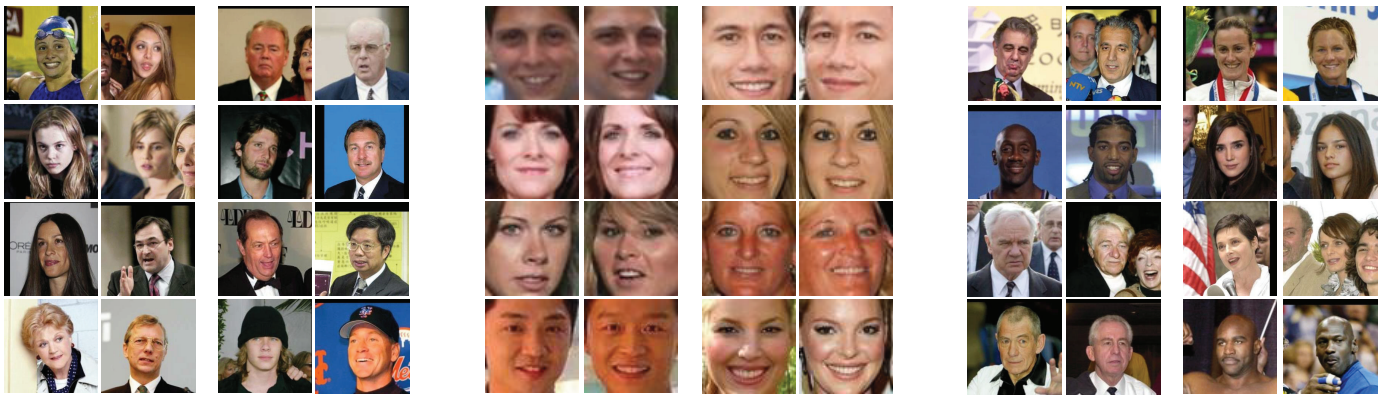

Figure 2. Negative pairs of face images sampled from the LFW, FGFV and FGLFW datasets. (**a**) LFW, (**b**) FGFV, (**c**) FGLFW.

4.1.2. Experimental Results

This section evaluates the proposed MVCSL method and the baseline methods with several traditional metric learning methods, namely ITML [18], side-information-based linear discriminant analysis (SILD) [29], KISSME [10], similarity metric learning over intra-personal subspace (Sub-SML) [30] and CSML [19]. Tables 1 and 2 show the mean accuracies and standard error of various methods under restricted settings on the FGFV and FGLFW datasets, respectively. The ITML, SILD and KISSME methods are formulated under the Mahalanobis distance framework, while CSML and MVCSL are designed under the cosine similarity framework. Compared with Mahalanobis distance-based methods, cosine similarity-based methods achieve better performance, and our proposed MVCSL obtains the best performance on both FGFV and FGLFW datasets. The reason is that our MVCSL can collaboratively learn multiple similarity measures from multiple feature representations to supplement each other with the difference information. In addition, Figures 3 and 4 plot the receiver operating characteristic (ROC) curves of various approaches on the FGFV and FGLFW datasets, respectively, and these experiments further show the promising performance of the proposed MVCSL method.

Table 1. Mean accuracy and standard error (%) of different methods on the FGFV dataset under the image restricted setting.

Method	**HOG**	**LBP**	**SIFT**
ITML	63.52 ± 4.41	62.86 ± 3.84	64.29 ± 4.29
KISSME	69.67 ± 3.37	68.35 ± 3.26	69.67 ± 3.40
SILD	70.00 ± 3.68	62.53 ± 3.17	68.57 ± 3.53
CSML	71.43 ± 1.94	72.31 ± 3.53	72.31 ± 3.24
MVC-s	79.23 ± 3.35	81.43 ± 1.96	78.90 ± 3.49
Method		**HOG, LBP, SIFT**	
Con		84.95 ± 2.29	
MVC-i		82.31 ± 2.73	
MVCSL		86.70 ± 2.62	

4.2. Kinship Verification

Kinship verification aims to predict whether a given pair of face images has a kind of kin relationship or not, which is a challenging subtask of face verification and has attracted a lot of attention in pattern recognition and computer vision.

Table 2. Mean accuracy and standard error (%) of different methods on the FGLFW dataset under the image restricted setting.

Method	HOG	LBP	SIFT
ITML	64.48 ± 1.54	65.07 ± 1.74	62.32 ± 1.84
KISSME	65.43 ± 1.29	66.60 ± 2.04	63.17 ± 2.36
Sub-SML	67.88 ± 2.32	69.18 ± 0.78	65.83 ± 2.01
CSML	68.00 ± 2.30	68.98 ± 2.83	67.87 ± 1.67
MVC-s	70.52 ± 2.22	71.18 ± 2.89	70.00 ± 1.39
Method	**HOG, LBP, SIFT**		
Con	71.62 ± 1.51		
MVC-i	73.33 ± 2.40		
MVCSL	74.23 ± 2.14		

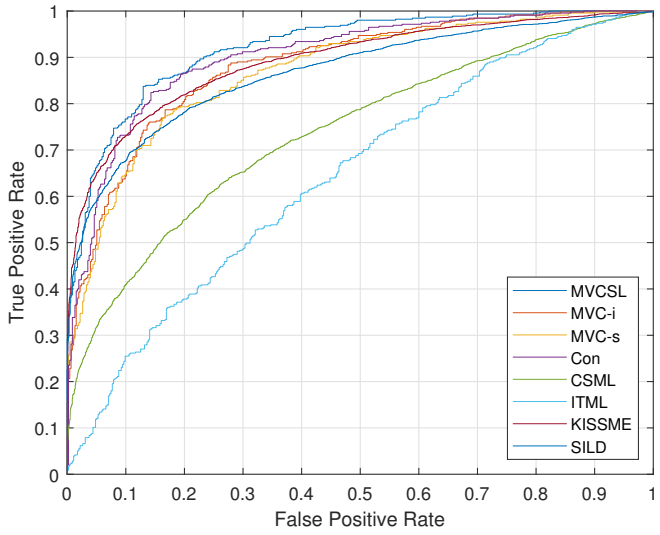

Figure 3. ROC curves of various approaches on the FGFV dataset.

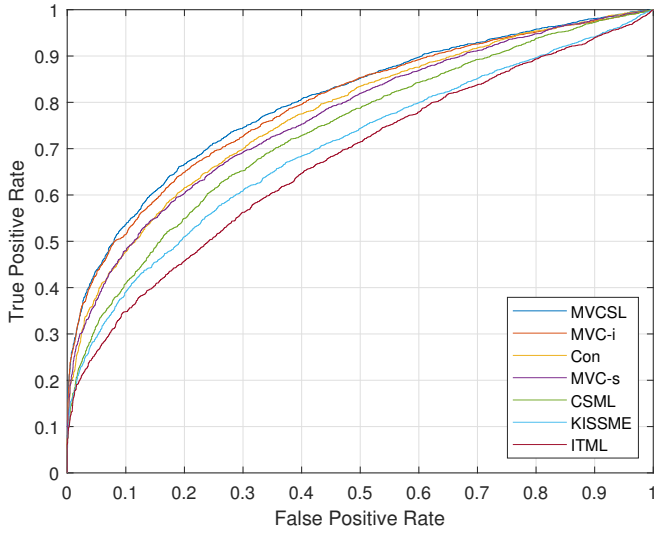

Figure 4. ROC curves of various approaches on the FGLFW dataset.

4.2.1. Dataset and Settings

In this subsection, we evaluate the proposed MVCSL approach in KinFaceW-I [7] and KinFaceW-II [7] datasets for the kinship verification task. The samples of them were collected from the unconstrained conditions with the obvious variations on lighting, age,

expression and posture. There are four kin relationships in them, i.e., father son (F-S), father daughter (F-D), mother son (M-S) and mother daughter (M-D).

Referring to the experimental settings provided by the datasets, the positive pair consists of two face images with kin relationship, and each negative pair is randomly selected from two unrelated face images without kinship. In order to reduce the background information of the sample image, we use the aligned KinFaceW-I and KinFaceW-II datasets, where each image was scaled to the size of 64×64 pixels. For feature representation, we adopt the same setting as the FGFV dataset and extract LBP, HOG and SIFT features for each sample, and each feature representation is reduced to 200 dimensions by PCA. According to the benchmark protocol of the KinFaceW-I and KinFaceW-II datasets, we adopt the positive and negative samples under the image restricted setting. In the experiment, we use a 5-fold cross-evaluation strategy to divide the positive and negative sample pairs into five groups, four groups for training and one group for test.

4.2.2. Experimental Results

We evaluate the MVCSL method with three evaluation strategies in the KinFaceW-I and KinFaceW-II datasets. Tables 3 and 4 list the mean verification accuracy (%) on two datasets. The mean verification accuracies of our MVC-s method with LBP, HOG and SIFT descriptors are 73.15%, 77.07% and 75.91% on the KinFaceW-I dataset, and those are 75.50%, 79.50% and 80.05% on the KinFaceW-II dataset, respectively. We also notice that the MVC-i and Con methods can learn the discriminative information of multiple feature representations, and our MVCSL further mines the potential information between various features. In Tables 3 and 4, we also provide the comparisons of the proposed MVCSL and several representative kinship verification methods, and these compared methods include block-based neighborhood repulsed metric learning (BNRML) [31], geometric mean metric learning (GMML) [32], multi-view geometric mean metric learning (MVGMML) [33], discriminative compact binary face descriptor (D-CBFD) [34], local large-margin multi-metric learning (L^2M^3L) [25], and weakly supervised compositional metric learning (WSCML) [35]. We can see from two tables that our MVCSL method obtains a competitive performance on both KinFaceW-I and KinFaceW-II datasets. Moreover, Figures 5 and 6 plot the ROC curves of the proposed MVCSL and baseline approaches on two kinship datasets, respectively. Experimental results on the benchmark kinship verification datasets intuitively show that our MVCML approach is able to efficiently exploit the common information of different feature representations and the private information of each feature representation to help improve the performance of kinship verification task.

Table 3. Comparisons of mean verification accuracy (%) on the KinFaceW-I dataset under the image-restricted setting.

Method	F-S	F-D	M-S	M-D	Mean
MVC-s (LBP)	77.57	70.17	68.04	76.84	73.15
MVC-s (HOG)	82.37	73.53	73.22	79.16	77.07
MVC-s (SIFT)	81.42	74.27	71.52	76.41	75.91
MVC-i	82.69	73.53	71.97	80.36	77.14
Con	83.01	74.64	72.81	79.96	77.61
MVCSL	84.30	75.38	74.53	81.16	78.84
BNRML [31]	76.28	70.51	73.70	72.47	73.24
GMML [33]	69.28	72.42	69.42	74.36	71.37
MVGMML [33]	69.25	75.00	69.40	72.76	71.13
D-CBFD [34]	79.60	73.60	76.10	81.50	77.60
WSCML [35]	81.90	73.95	72.88	72.90	75.21

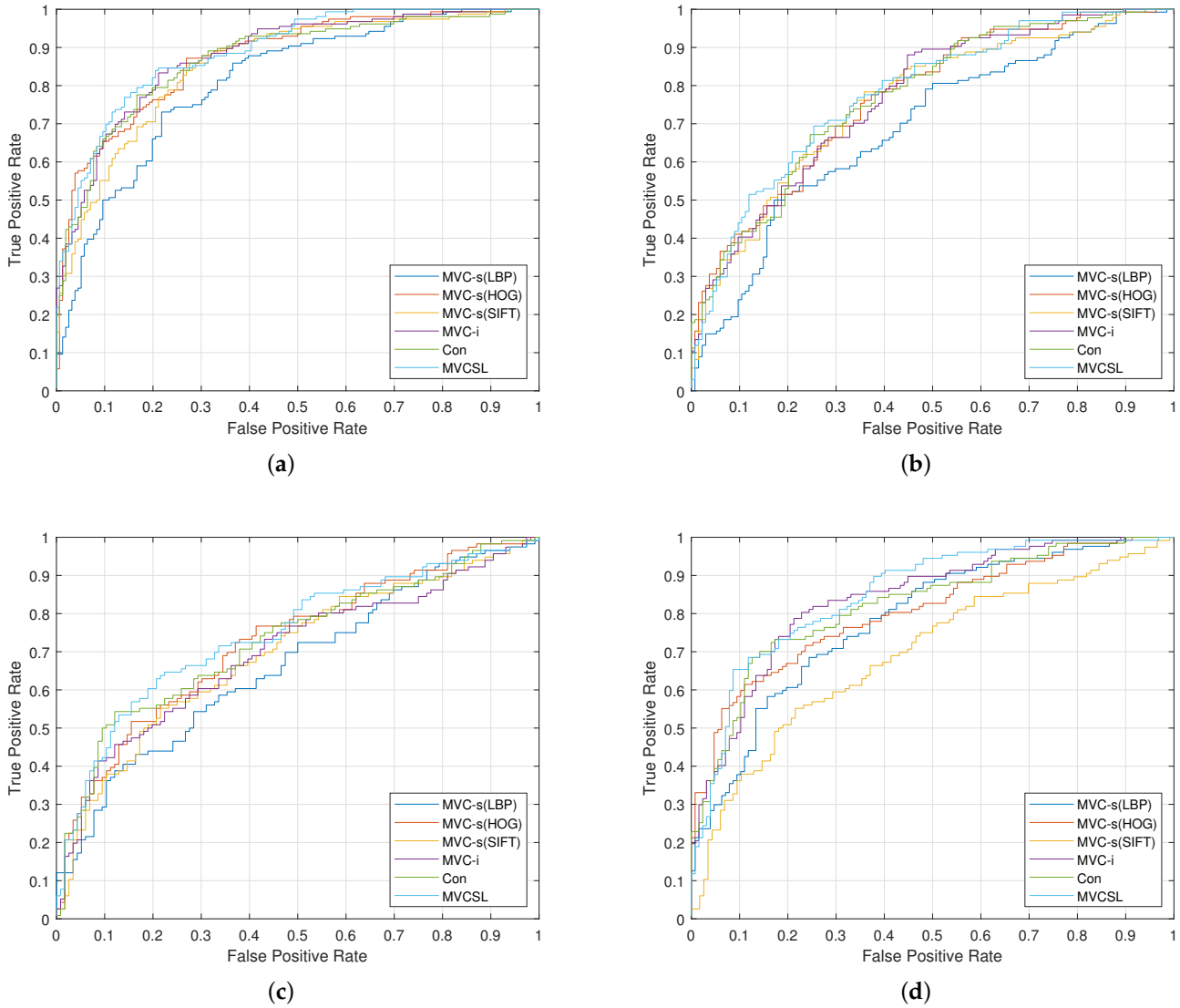

Figure 5. ROC curves of the proposed MVCSL and baseline methods on the KinFaceW-I dataset. (**a**) F-S, (**b**) F-D, (**c**) M-S, (**d**) M-D.

Table 4. Comparisons of mean verification accuracy (%) on the KinFaceW-II dataset under the image-restricted setting.

Method	F-S	F-D	M-S	M-D	Mean
MVC-s (LBP)	78.80	76.80	74.60	71.80	75.50
MVC-s (HOG)	83.80	76.40	79.60	76.40	79.05
MVC-s (SIFT)	83.00	77.60	81.00	78.60	80.05
MVC-i	83.80	77.60	81.20	78.80	80.35
Con	83.60	78.03	81.00	78.00	80.15
MVCSL	84.80	79.00	81.80	78.40	81.00
BNRML [31]	79.40	79.00	77.00	72.80	77.05
GMML [33]	68.60	73.20	67.80	68.40	69.50
MVGMML [33]	70.40	73.40	65.80	69.20	69.70
D-CBFD (HOG) [34]	81.00	76.20	77.40	79.30	78.50
L^2M^3L [25]	82.40	78.20	78.80	80.40	80.00

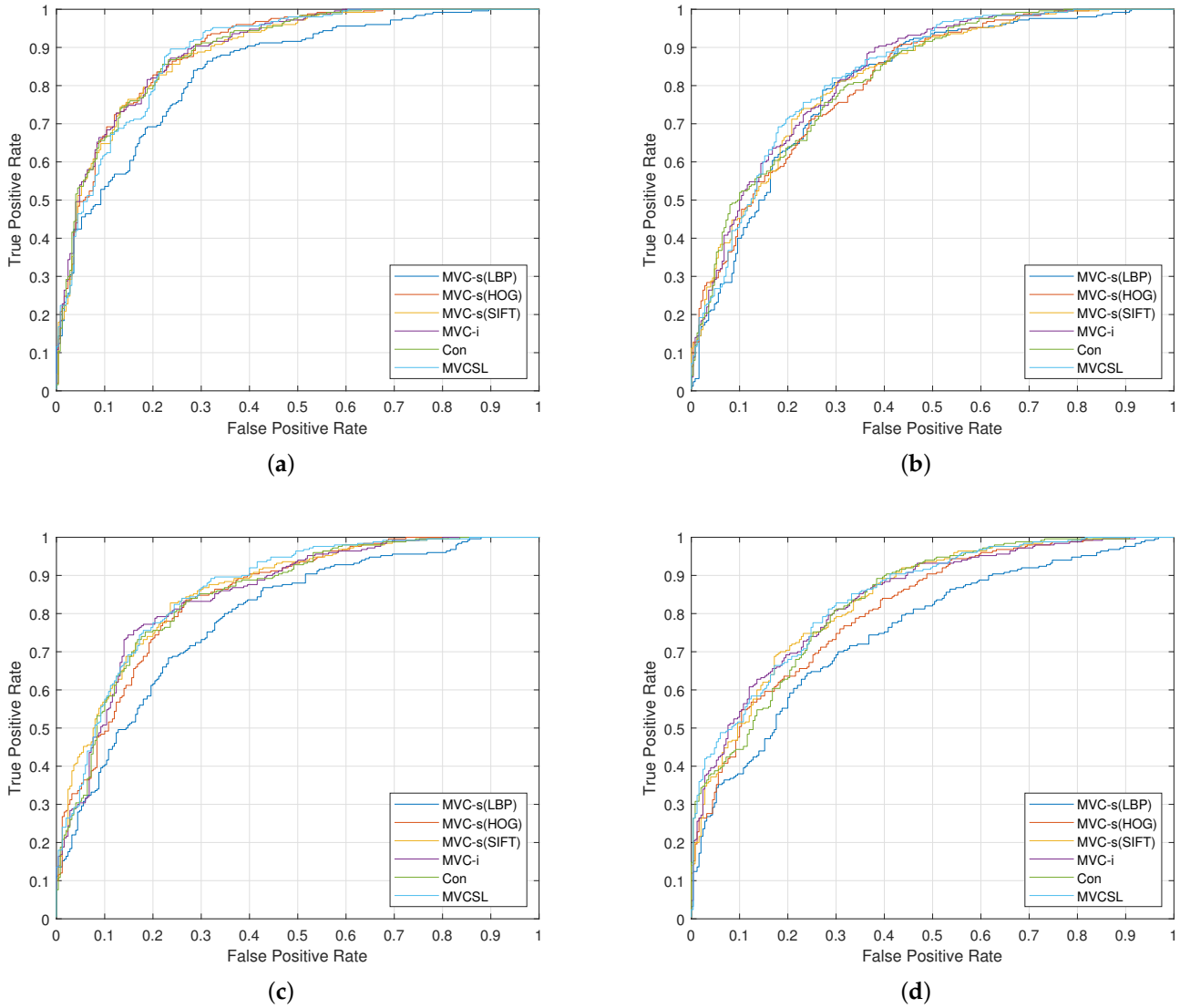

Figure 6. ROC curves of the proposed MVCSL and baseline methods on the KinFaceW-II dataset. (**a**) F-S, (**b**) F-D, (**c**) M-S, (**d**) M-D.

5. Conclusions

This paper proposes a multi-view cosine similarity learning (MVCSL) approach to make use of multi-view feature representations of data and apply it for fine-grained face verification and facial kinship verification tasks. The proposed MVCSL method can complement each other with the difference information among multiple views by jointly learning a cosine similarity for each view in a unified framework. Moreover, in order to mine non-trivial samples, we set the margin to make sure that the joint cosine similarity of positive pairs is greater than a large value and the joint cosine similarity of negative pairs is less than a small value. Experimental results on the fine-grained face verification and facial kinship verification tasks demonstrate the advantages of our MVCSL method for exploiting multi-view data.

The main novelty and contribution of our work is to advance the multi-view metric learning from the cosine similarity learning framework to better exploit multi-view data, which is different from the existing multi-view metric learning methods that are mainly formulated in the framework of Mahalanbis distance metric learning. The shortcoming of the proposed MVCSL method is that the gradient-descent based method is used to find the

linear transformation matrices, and we may not obtain the global optimal solution. In the future, we hope we can further improve our MVCSL and achieve its closed-form solution. In future work, we will apply our approach to other applications such as visual recognition, classification and clustering in pattern recognition.

Author Contributions: Conceptualization, J.H.; methodology, Z.W. and J.C.; software, Z.W. and J.C.; validation, Z.W., J.C. and J.H.; formal analysis, Z.W., J.C. and J.H.; investigation, Z.W. and J.C.; resources, J.H.; data curation, J.H.; writing—original draft preparation, Z.W. and J.C.; writing—review and editing, J.H.; visualization, J.C.; supervision, J.H.; project administration, J.H.; funding acquisition, J.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under grant number 62006013.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Suárez, J.L.; García, S.; Herrera, F. A tutorial on distance metric learning: Mathematical foundations, algorithms, experimental analysis, prospects and challenges. *Neurocomputing* **2021**, *425*, 300–322. [CrossRef]
2. Schroff, F.; Kalenichenko, D.; Philbin, J. Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 815–823.
3. Zheng, W.; Lu, J.; Zhou, J. Hardness-Aware Deep Metric Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 3214–3228. [CrossRef] [PubMed]
4. Karlinsky, L.; Shtok, J.; Harary, S.; Schwartz, E.; Aides, A.; Feris, R.; Giryes, R.; Bronstein, A.M. Repmet: Representative-based metric learning for classification and few-shot object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5197–5206.
5. 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: Amherst, MA, USA, 2007.
6. Kemelmacher-Shlizerman, I.; Seitz, S.M.; Miller, D.; Brossard, E. The megaface benchmark: 1 million faces for recognition at scale. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4873–4882.
7. Lu, J.; Zhou, X.; Tan, Y.; Shang, Y.; Zhou, J. Neighborhood Repulsed Metric Learning for Kinship Verification. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 331–345. [PubMed]
8. Wang, M.; Deng, W. Deep face recognition: A survey. *Neurocomputing* **2021**, *429*, 215–244. [CrossRef]
9. Guillaumin, M.; Verbeek, J.; Schmid, C. Is that you? Metric learning approaches for face identification. In Proceedings of the IEEE International Conference on Computer Vision, Kyoto, Japan, 27 September–4 October 2009; pp. 498–505.
10. Koestinger, M.; Hirzer, M.; Wohlhart, P.; Roth, P.M.; Bischof, H. Large scale metric learning from equivalence constraints. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 2288–2295.
11. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
12. Ahonen, T.; Hadid, A.; Pietikainen, M. Face description with local binary patterns: Application to face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 2037–2041. [CrossRef] [PubMed]
13. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 886–893.
14. Chen, N.; Zhu, J.; Sun, F.; Xing, E.P. Large-margin predictive latent subspace learning for multiview data analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2365–2378. [CrossRef] [PubMed]
15. Hu, J.; Lu, J.; Tan, Y.P. Sharable and individual multi-view metric learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 2281–2288. [CrossRef] [PubMed]
16. Xing, E.P.; Jordan, M.I.; Russell, S.J.; Ng, A.Y. Distance metric learning with application to clustering with side-information. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 9–14 December 2002; pp. 521–528.
17. Weinberger, K.Q.; Saul, L. Distance Metric Learning for Large Margin Nearest Neighbor Classification. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 4–7 December 2005; pp. 1473–1480.
18. Davis, J.V.; Kulis, B.; Jain, P.; Sra, S.; Dhillon, I.S. Information-theoretic metric learning. In Proceedings of the Twenty-Fourth International Conference on Machine Learning, Corvallis, OR, USA, 20–24 June 2007; pp. 209–216.

19. Nguyen, H.V.; Bai, L. Cosine similarity metric learning for face verification. In Proceedings of the 10th Asian Conference on Computer Vision, Queenstown, New Zealand, 8–12 November 2010; pp. 709–720.
20. Tang, Z.; Wu, X.; Fu, B.; Chen, W.; Feng, H. Fast face recognition based on fractal theory. *Appl. Math. Comput.* **2018**, *321*, 721–730. [CrossRef]
21. Gdawiec, K.; Domanska, D. Partitioned iterated function systems with division and a fractal dependence graph in recognition of 2D shapes. *Int. J. Appl. Math. Comput. Sci.* **2011**, *21*, 757–767. [CrossRef]
22. Tan, T.; Yan, H. Face recognition using the weighted fractal neighbor distance. *IEEE Trans. Syst. Man, Cybern. Part C (Appl. Rev.)* **2005**, *35*, 576–582. [CrossRef]
23. Wang, X.; Han, X.; Huang, W.; Dong, D.; Scott, M.R. Multi-similarity loss with general pair weighting for deep metric learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 5022–5030.
24. Xie, P.; Xing, E. Multi-Modal Distance Metric Learning. In Proceedings of the International Joint Conference on Artificial Intelligence, Beijing, China, 3–9 August 2013; pp. 1806–1812.
25. Hu, J.; Lu, J.; Tan, Y.P.; Yuan, J.; Zhou, J. Local large-margin multi-metric learning for face and kinship verification. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 1875–1891. [CrossRef]
26. Jia, X.; Jing, X.Y.; Zhu, X.; Chen, S.; Du, B.; Cai, Z.; He, Z.; Yue, D. Semi-supervised Multi-view Deep Discriminant Representation Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *43*, 2496–2509. [CrossRef] [PubMed]
27. Hu, J.; Lu, J.; Tan, Y.P. Fine-grained face verification: Dataset and baseline results. In Proceedings of the International Conference on Biometrics, Phuket, Thailand, 19–22 May 2015; pp. 79–84.
28. Deng, W.; Hu, J.; Zhang, N.; Chen, B.; Guo, J. Fine-grained face verification: FGLFW database, baselines, and human-DCMN partnership. *Pattern Recognit.* **2017**, *66*, 63–73. [CrossRef]
29. Kan, M.; Shan, S.; Xu, D.; Chen, X. Side-Information based Linear Discriminant Analysis for Face Recognition. In Proceedings of the British Machine Vision Conference, Dundee, Scotland, 29 August–2 September 2011; pp. 1–12.
30. Cao, Q.; Ying, Y.; Li, P. Similarity metric learning for face recognition. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 2408–2415.
31. Patel, B.; Maheshwari, R.; Raman, B. Evaluation of periocular features for kinship verification in the wild. *Comput. Vis. Image Underst.* **2017**, *160*, 24–35. [CrossRef]
32. Zadeh, P.; Hosseini, R.; Sra, S. Geometric Mean Metric Learning. In Proceedings of the 33nd International Conference on Machine Learning, New York, NY, USA, 19–24 June 2016; pp. 2464–2471.
33. Hu, J.; Lu, J.; Liu, L.; Zhou, J. Multi-view geometric mean metric learning for kinship verification. In Proceedings of the IEEE International Conference on Image Processing, Taipei, China, 22–25 September 2019; pp. 1178–1182.
34. Yan, H. Learning discriminative compact binary face descriptor for kinship verification. *Pattern Recognit. Lett.* **2019**, *117*, 146–152. [CrossRef]
35. Chen, J.; Hu, J. Weakly Supervised Compositional Metric Learning for Face Verification. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 1–8. [CrossRef]