Fast High Dimensional Vector Multiplication Face Recognition

Oren Barkan
Tel Aviv University
orenbarkan@post.tau.ac.il

Jonathan Weill
Tel Aviv University
yonathanw@post.tau.ac.il

Lior Wolf
Tel Aviv University
wolf@cs.tau.ac.il

Hagai Aronowitz
IBM Research
hagaia@il.ibm.com

Abstract

This paper advances descriptor-based face recognition by suggesting a novel usage of descriptors to form an over-complete representation, and by proposing a new metric learning pipeline within the same/not-same framework. First, the Over-Complete Local Binary Patterns (OCLBP) face representation scheme is introduced as a multi-scale modified version of the Local Binary Patterns (LBP) scheme. Second, we propose an efficient matrix-vector multiplication-based recognition system. The system is based on Linear Discriminant Analysis (LDA) coupled with Within Class Covariance Normalization (WCCN). This is further extended to the unsupervised case by proposing an unsupervised variant of WCCN. Lastly, we introduce Diffusion Maps (DM) for non-linear dimensionality reduction as an alternative to the Whitened Principal Component Analysis (WPCA) method which is often used in face recognition.

We evaluate the proposed framework on the LFW face recognition dataset under the restricted, unrestricted and unsupervised protocols. In all three cases we achieve very competitive results.

1. Introduction

The Labeled Faces in the Wild (LFW) face recognition benchmark [1] is currently the most active research benchmark of its kind. It is built around a simple binary decision task: given two face images, is the same person being photographed in both? The comprehensive results tables show a large variety of methods which can be roughly divided into two categories: pair comparison methods and signature based methods.

In the pair comparison methods [2, 3, 4], the decision is based on a process of comparing the two images part by part, oftentimes involving an iterative local matching process. In the signature based methods [5, 6, 7, 8, 9], each face image is represented by a single descriptor vector and is then discarded. To compare two face images, their signatures are compared using predefined metric functions, which are sometimes learned based on the training data.

The pair comparison methods allow for flexibility in representation, based on the actual image pair to be compared. On the other hand, the signature based methods are often much more efficient. Furthermore, there is a practical value in signature based methods in which the signature is compact. Such systems can store and retrieve face images using limited resources.

In this paper, we propose an efficient signature based method, in which the storage footprint of each signature is on the order of a hundred floating point numbers. This compares to storage footprints of one to three orders of magnitude larger in previous work.

Our method includes multiple contributions. First, as detailed in Section 2, we propose to use over-complete representations of the input image. This is shown to significantly contribute to the overall performance. However, this added accuracy is hidden until dimensionality reduction is performed. In Section 3, we propose the use of the WCCN [10] metric learning technique for face recognition. In Section 4, we propose a general scheme for generating labeled data from an unlabeled data. In Section 5, we describe in detail our proposed recognition system, which is applicable for both supervised and unsupervised learning by utilizing the scheme described in Section 4. This results in an extension of the WCCN metric learning to the unsupervised case. In Section 6, the Diffusion Maps technique (DM) [11] is introduced as a non-linear dimensionality reduction method for face recognition. We investigate it as an alternative to WPCA [6] and show that it can improve performance over the baseline when being fused with WPCA. In Section 7, we evaluate the proposed system on the LFW dataset under the restricted, unrestricted and unsupervised protocols and report state of the art results on these benchmarks. Finally, in Section 8, we conclude and discuss future work.

1.1 Overview of the recognition pipeline

A unified pipeline is used in order to solve the unsupervised case and the two supervised scenarios of the LFW benchmark: the restricted and the unrestricted protocols.

First, a representation is constructed from the face images. This either uses existing methods, such as LBP [9], TPLBP [12] and SIFT [7], or methods which are introduced to the fields in this paper, such as the OCLBP

and the use of the Scattering transform [13]. Second, a dimensionality reduction step takes place. This is either WPCA or the Diffusion Maps for the unsupervised case, or PCA-LDA or DM-LDA for the two supervised settings. Third, WCCN is applied. For the supervised settings, the original WCCN method [10] is applied. For the unsupervised case, our unsupervised WCCN variant is applied. As a last step, cosine similarities based on multiple representations and image features are combined together using uniform weighting.

2. Over-complete representations

Over-complete representations have been found to be useful for improving the robustness of classification systems by using richer descriptors [14, 15]. In this work, we introduce two new adaptations of descriptors for the domain of face recognition. Both of them share the property of over-complete representation. In the experimental results section, we show that the improvement in the accuracy of using over-complete representations remains hidden until some dimensionality reduction is involved. However, its contribution to the final score is significant.

2.1. Over-complete local binary patterns

LBP [16] is one of the most successful features for texture classification. Specifically, a modified 'uniform' version [9] of the original LBP was found to be useful for the task of face recognition. Several attempts to extend or modify the LBP have been made in [12, 17]. However, most of them resulted in new variants of LBP which do not necessarily outperform the original one.

The standard LBP operator for face recognition is denoted as $LBP^{u2}_{p,r}$ where $u2$ stands for uniform patterns, p defines the number of points that are uniformly sampled over a circle with a radius r. This computation is done block-wise and the results from all blocks are concatenated to form a final descriptor. For an overview of the LBP operator for face recognition we refer the reader to [9].

In this work we keep the original form of the LBP as it is, but suggest an over-complete representation built on top of it. The proposed Over-Complete LBP (OCLBP) differs from the original LBP in two major properties. First, it is computed with overlapping blocks, similar to [18]. The amount of vertical- and horizontal-overlap is controlled by the two parameters $v, h \in [0,1)$ with $h = v = 0$ degenerating to non-overlapping blocks. The second difference is in the varied block and radius sizes. We repeat the LBP computation for different sizes of block and radius, similar to the multi-scale variant in [19]. We name the resulting representation as OCLBP. More formally, given an input image and a set of configurations $S = \{(a_i, b_i, v_i, h_i, p_i, r_i)\}_{i=1}^{k}$, we divide the image to blocks in a size of $a_i \times b_i$ with vertical overlap of v_i, horizontal overlap of h_i and compute a LBP descriptor using the operator $LBP^{u2}_{p_i,r_i}$. We repeat this computation for all configurations in S and concatenate the descriptors to a single vector which is the resulting OCLBP descriptor.

Since the computations of the different configurations are independent, the OCLBP descriptor can be easily paralleled.

We show in Section 7 that the OCLBP descriptor achieves the same performance as the standard LBP when they are used in their original dimension. However, after applying dimensionality reduction, a significant gain in accuracy is achieved by the more elaborate scheme.

2.2. Scattering transform for face recognition

The Scattering Transform was introduced by Mallat in [13]. This work has been extended to various computer vision tasks in [20, 21]. As an image representation, a scattering convolution network was proposed in [20]. This representation leads to an extremely high dimensional descriptor that is invariant for small local deformations in the image. For texture classification, a Scattering wavelet network managed to achieve state of the art results [21].

The output of the first layer of a scattering network can be considered as a SIFT-like descriptor while the second layer adds further complementary invariant information which improves discrimination quality. The third layer, however, was found to have a negligible contribution for classification accuracy while increasing the computational cost significantly.

In this work, we investigate the contribution of the Scattering descriptor to our face recognition framework. In a similar manner to the OCLBP, we find that the Scattering descriptor is much more effective when combined with dimensionality reduction.

We refer the reader to [13] for a detailed description of the Scattering transform.

3. Within class covariance normalization

Within Class Covariance Normalization (WCCN) has been used mostly in the speaker recognition community and was first introduced in [10]. The within class covariance matrix W is computed as follows:

$$W = \frac{1}{C}\sum_{i=1}^{C}\frac{1}{n_i}\sum_{j=1}^{n_i}(x_i^j - \mu_i)(x_i^j - \mu_i)^T,$$

Where C is the number of different classes, n_i is the number of instances belonging to class i, x_i^j is the jth instance of class i and μ_i is the mean of class i.

In a sense, WCCN is similar to the family of methods

that down-regulate the contribution of the directions in the vector space that account for much of the within class covariance. This is often done by projecting the data onto the subspace spanned by the eigenvectors corresponding to the smallest eigenvalues of W. In WCCN, this effect is performed in a softer way without performing explicit dimensionality reduction: instead of discarding the directions that correspond to the top eigenvalues, WCCN reduces the effect of the within class directions by employing a normalization transform $T = W^{-1/2}$. While to the best of our knowledge it was previously unused in face recognition, we show a clear improvement in performance over the state of the art by using the WCCN method when applied in the LDA subspace.

In this work, we also introduce an unsupervised version of WCCN, which is shown to be useful in case we lack the necessary labeled data. In Section 7, we evaluate our proposed method and show that it is an improvement over the baseline algorithms. Furthermore, we show that although the unsupervised WCCN algorithm does not make use of any label information, it is competitive with the original supervised WCCN in several scenarios.

4. Unsupervised labeling

A common and challenging problem in machine learning is the beneficial utilization of successful supervised algorithms in the absence of labeled data. In this section, we propose a simple unsupervised algorithm for generating valuable labels for the pair matching problem.

Before describing the algorithm, we enumerate our two assumptions. First, we assume that we are equipped with an unsupervised algorithm that is able to achieve some classification accuracy – we consider this algorithm as the baseline algorithm. We focus our discussion on algorithms that produce a classification score and not just binary labels. The second assumption is on the shape of the distribution of the classification scores. We assume that the score distribution is approximately uni-modal and has two tails. If our baseline algorithm manages to achieve a reasonable accuracy on the training set, we would expect to find many fewer classification mistakes on the tails, rather than in the area around the mean score.

Algorithm 1 (A, B, T, t_l, t_r)

Inputs: *A - a trained model of the baseline unsupervised algorithm, B - supervised algorithm, T - training set. t_l - a threshold on the left tail, t_r - a threshold on the right tail.*

Output: *C - a new trained model.*

1. *Compute the pair-wise score matrix S using A and T.*
2. *Assign a label of 1 to all pairs with a score above t_r.*
3. *Assign a label of -1 to all pairs with a score below t_l.*
4. *Assign a label of 0 to all the other pairs.*
5. *Train a new model C using the assigned labels and B*
6. **Return** *C*

In the case of the "same/not-same" classification, we would expect the majority of the scores in one tail to belong to pairs that are matched and the majority of the scores on the other tail to belong to pairs that are mismatched. This behavior leads to the formation of two (hopefully) separated sets: one consists mostly of "same" pairs and the other consists mostly of "not-same" pairs. The size of each cluster is determined by the number of pairs we pick from the corresponding tail. This number is a parameter that defines a tradeoff between the number of desired labels and the confidence that we have in this labeling. Therefore, we propose Algorithm 1.

Note that except for positive and negative labels there are also 'unknown' labels. In case we are equipped with an algorithm (B) that is designed to handle unlabeled samples (i.e., a semi-supervised algorithm), we provide it with this information. Otherwise, we provide B exclusively with the positive and negative sets of examples.

The optimal values of the parameters t_l and t_r are related to the accuracy of the baseline model A, the shape of the score distribution, and the number of labels that we want to generate. For example, if we are provided with a baseline model which achieves poor accuracy, we should expect poor labeling as well. In case the empirical distribution is symmetric we can choose $t_l = t_r$, otherwise we might consider the size of the tails for each tail separately. Since the generated labels are used to train a new supervised model we can apply Algorithm 1 iteratively. Another possible extension is to use a set of supervised algorithms instead of a single one and to determine the final labeling according to a voting scheme.

5. Fast supervised and unsupervised vector multiplication recognition system

We now describe in detail our proposed recognition system, which we call VMRS for Vector Multiplication Recognition System. Given two samples, we need to decide whether they belong to the same class or not. First, each sample is projected to a low dimensional subspace by WPCA. Then, we perform an additional supervised dimensionality reduction by applying LDA. Finally, we perform WCCN to the resultant feature vectors in the low dimensional LDA-subspace and produce a score by applying cosine similarity. Therefore, the pipeline can be reduced to two matrix-vector multiplications followed by cosine similarity. We formally denote P, L and W as the

WPCA projection matrix, LDA projection matrix and Within Class Covariance (WCC) matrix, respectively. Thus, given two vectors, $x, y \epsilon R^n$, representing two face images, the final score is defined as:

$$s(x, y, M) = \frac{(Mx)^T (My)}{\left|Mx\right|\left|My\right|}$$

Where, $M = W^{-1/2} LP$.

The final decision is made according to a prescribed threshold that can be set to an Equal Error Rate (EER) point, Verification Rate (VR) point, or alternatively, can be learned by a SVM [22].

5.1. Unsupervised pipeline

The pipeline described above is supervised and requires labeled data. However, in many real-world scenarios we lack labels. In such cases we can apply Algorithm 1 (Section 4) in order to generate artificial labels for the training set. Specifically, we use the WPCA model as a baseline A and generate new labels according to the distribution of the scores of pairs in the training set. We then use these labels to estimate the within class covariance matrix (note that we do not apply LDA in this case, since it is unsupervised). Since WCCN computation is based on pairs from the same class, we only choose scores from one of the two tails (the 'same' tail). Then we treat each pair in the 'same' group as a single class and merge classes that share the same samples, i.e., we utilize strongly connected components in the connectivity graph induced by the similar pairs.

In our experiments, we selected the parameter t_l so that the pairs with distances in the bottom 15% of the distances of all possible pairs will constitute the "same" pairs. This value was determined once, when performing a limited investigation of View 1 of the LFW benchmark (intended for parameter fitting) and remained fixed. In Section 7, we show that this approach improves over the baseline WPCA system.

As already mentioned in Section 4, one can iterate between generating new labels, using them for training a new supervised model, and generating new scores. However, we did not find that performing multiple iterations improves performance. Hence, Algorithm 1 is employed only once. With the introduction of this unsupervised variant of WCCN, the proposed system is suitable for both the supervised and the unsupervised scenarios.

It is important to clarify that our proposed system, excluding the feature extraction phase, is extremely efficient in the sense of computational complexity. The most demanding computation which takes place during the test phase is the linear transformation M on the pair of original feature vectors x, y. This has a great advantage over "lazy" learning approaches such as [22] which make an explicit use of the training set during the test phase. The complexity of the training phase is dominated by the complexity of the computation of the eigen-problems that are encountered in WPCA and LDA and the computation of the matrix square root of W^{-1}.

6. Diffusion Maps

Many of the state of the art face recognition systems incorporate a dimensionality reduction component. The aim of dimensionality reduction is twofold. First, learning in high dimensional vector spaces is computationally demanding. Second, in some cases and especially when the high dimensionality stems from over-complete representations, there is a large amount of redundancy in the data. Dimensionality reduction techniques attempt to solve both of these problems by exploring meaningful connections between the data points and discover the geometry that best represents that data. Most of the work done so far in face recognition applied linear dimensionality reduction. One of the problems with linear dimensionality reduction is the implicit assumption that the geometric structure of data points is well captured by a linear subspace. It has been shown [23] that real world signals, in most cases, have non-linear structures and reside over a manifold.

We propose to use a non-linear dimensionality reduction technique called Diffusion Maps (DM). We introduce a whitened variant of the conventional DM framework and show how to deal with the out-of-sample extension problem, which occurs in the test phase. In Section 7, we show that by incorporating the DM framework into the proposed recognition system of Section 5, we achieve results which are on a par with the state of the art. Finally, we show that by combining DM and WPCA we are able to get an additional improvement in accuracy.

We will briefly describe the main steps of DM (for a fully rigorous mathematical derivation we refer the reader to [11]).

In the DM framework, we are provided with a training set $\{x_i\}_{i=1}^n \subset R^m$ and affinity kernel $k(\cdot,\cdot)$. A commonly used kernel is the Gaussian kernel:

$$k(x_i, x_j) = \exp\left(-\frac{c(x_i, x_j)^2}{\sigma}\right)$$

Where $c(\cdot,\cdot)$ is a metric and σ is a parameter which determines the size of the neighborhood over which we trust our local similarity measure. Using the affinity kernel, we compute a pair-wise affinity matrix K. Then, we convert K to a transition Markov matrix P by normalizing each row in K by its sum: $P = D^{-1}K$, where D is a diagonal matrix normalizing the rows of K. Therefore, P^t is a matrix, in which the entry $P_{i,j}^t$ is the

Tables 1-3: Classification accuracy (± standard error) of various combinations of classifiers and descriptors in the unsupervised, restricted and unrestricted settings, respectively. See text for details regarding the classifiers and descriptors.

Table 1	LBP		OCLBP		TPLBP		SIFT		SCATTERING	
Unsupervised		SQRT		SQRT		SQRT		SQRT		SQRT
RAW	72.48 ± 0.49	72.48 ± 0.49	72.78 ± 0.39	72.78 ± 0.39	73.91 ± 0.57	73.91 ± 0.57	68.43 ± 0.49	68.43 ± 0.49	66.83 ± 0.63	66.83 ± 0.63
WPCA	77.90 ± 0.59	80.55 ± 0.38	80.21 ± 0.35	82.78 ± 0.41	78.06 ± 0.45	79.71 ± 0.48	78.80 ± 0.32	79.43 ± 0.30	80.01 ± 0.50	80.61 ± 0.48
DM	77.30 ± 0.60	79.56 ± 0.44	79.26 ± 0.42	82.20 ± 0.49	77.56 ± 0.40	78.55 ± 0.62	77.75 ± 0.33	78.96 ± 0.40	79.37 ± 0.56	81.13 ± 0.52
WPCA+WCCN	78.81 ± 0.73	82.48 ± 0.35	81.90 ± 0.42	86.66 ± 0.30	78.35 ± 0.52	80.2 ± 0.51	80.96 ± 0.43	81.88 ± 0.36	81.78 ± 0.49	82.50 ± 0.55
DM+WCCN	78.75 ± 0.58	82.43 ± 0.22	81.13 ± 0.40	85.46 ± 0.40	79.36 ± 0.43	81.33 ± 0.57	80.70 ± 0.35	81.91 ± 0.29	80.10 ± 0.55	81.36 ± 0.56

Table 2	LBP		OCLBP		TPLBP		SIFT		SCATTERING	
Restricted		SQRT		SQRT				SQRT		SQRT
PCALDA	83.30 ± 0.59	85.23 ± 0.37	85.10 ± 0.46	87.85 ± 0.69	82.71 ± 0.54	83.88 ± 0.62	83.30 ± 0.59	85.23 ± 0.37	85.10 ± 0.46	87.85 ± 0.69
DMLDA	81.53 ± 0.66	84.73 ± 0.50	84.68 ± 0.84	87.73 ± 0.58	80.13 ± 0.56	82.08 ± 0.62	81.53 ± 0.66	84.73 ± 0.50	84.68 ± 0.84	87.73 ± 0.58
WPCA	82.03 ± 0.59	84.86 ± 0.37	83.66 ± 0.50	87.23 ± 0.38	81.45 ± 0.61	82.91 ± 0.53	82.03 ± 0.59	84.86 ± 0.37	83.66 ± 0.50	87.23 ± 0.38
DM	81.91 ± 0.59	84.53 ± 0.33	83.76 ± 0.56	87.08 ± 0.33	80.05 ± 0.58	81.81 ± 0.59	81.91 ± 0.59	84.53 ± 0.33	83.76 ± 0.56	87.08 ± 0.33

Table 3	LBP		OCLBP		TPLBP		SIFT		SCATTERING	
Unrestricted		SQRT		SQRT		SQRT		SQRT		SQRT
PCALDA	84.40 ± 0.68	85.96 ± 0.58	86.78 ± 0.58	88.75 ± 0.59	83.91 ± 0.67	85.38 ± 0.67	86.61 ± 0.44	88.06 ± 0.19	87.00 ± 0.70	87.96 ± 0.70
DMLDA	83.23 ± 0.66	85.26 ± 0.59	85.71 ± 0.56	88.66 ± 0.60	82.91 ± 0.55	84.11 ± 0.59	86.80 ± 0.40	87.06 ± 0.36	85.88 ± 0.73	86.21 ± 0.73
WPCA	81.91 ± 0.63	84.53 ± 0.43	84.56 ± 0.45	87.30 ± 0.52	81.13 ± 0.70	83.31 ± 0.64	84.01 ± 0.58	84.85 ± 0.25	84.25 ± 0.60	84.89 ± 0.65
DM	81.11 ± 0.54	83.76 ± 0.48	83.61 ± 0.38	86.96 ± 0.53	81.58 ± 0.62	83.01 ± 0.58	82.93 ± 0.43	83.85 ± 0.34	83.87 ± 0.53	84.43 ± 0.62

probability of transition from node x_i to node x_j in t steps. A diffusion distance after t steps is defined by: $D_t(x_i, x_j) = \sum_{k=1}^{n} (P_{i,k}^t - P_{j,k}^t)^2$. Since the diffusion distance computation requires the evaluation of the distances over the entire training set, it results in an extremely complex operation. Fortunately, the same distance can be computed in a much simpler way: By spectral decomposition of P, we get a complete set of eigenvalues $1 = \lambda_0 \geq \lambda_1 \geq ... \geq \lambda_n$ and left and right eigenvectors satisfying: $P\psi_i = \lambda_i \varphi_i$. We then define a mapping $H_t : \{x_i\}_{i=1}^n \to V$ according to: $H_t(x_i) = \left[\lambda_1^t \psi_{1i}, ..., \lambda_l^t \psi_{li}\right]^T$, where ψ_{ki} indicates the i-th element of the k-th eigenvector of P and l is the dimension of the diffusion space V. It has been shown [11] that for $l = m-1$ the following equation holds: $\left\| H_t(x_i) - H_t(x_j) \right\|_2^2 = D_t(x_i, x_j)$. This result justifies the use of squared Euclidean distance in the diffusion space. In practice, one should pick $l < m-1$ according to the decay of $(\lambda_i)_{i=1}^n$. This decay is related to the complexity of the intrinsic dimensionality of the data and the choice of the parameter σ.

6.1. Uniform scaling

Inspired by WPCA, we propose to weigh all coordinates in the diffusion space uniformly. We do that by simply omitting the eigenvalues when computing the embedding. Therefore, we change the mapping H to $H(x_i) = \left[\psi_{1i}, ..., \psi_{li}\right]^T$.

While originally inspired by the relation between PCA and WPCA, this modification results in a significant improvement when applying it to DM framework. We hypothesize that this improvement occurs because DM, as an unsupervised algorithm, holds little information in its eigenvalues regarding the actual discrimination capability. Confounding factors, such as illumination, can be associated with some of the leading eigenvectors.

6.2. Out of sample extension

Since the domain of H is defined only on the training set, we cannot compute the embedding for a new test sample. A trivial solution would be to re-compute the spectral decomposition on the whole training data and the new test sample from scratch. However, this solution is extremely costly in the sense of computation time. Thus, we propose a simpler solution: Our approach assumes that the training data is sufficiently diverse in order to capture most of the variability of the face space. In this case, we would expect the embedding of a new test sample to be approximated well by a linear combination of embeddings of the training samples in the low dimensional diffusion space. A natural choice is to set the coefficient for each training sample as the probability of moving from it to the new test sample. Thus, for a new test sample x_{n+1} we compute the transition probabilities $P_{n+1,j}$, $\forall 1 \leq j \leq n$ and define its embedding to be $H(x_{n+1}) = \left[\sum_{j=1}^{n} P_{n+1,j}\psi_{1j}, ..., \sum_{j=1}^{n} P_{n+1,j}\psi_{lj}\right]^T$. As a result we get an extended mapping $H : \{x_i\}_{i=1}^{n+1} \to V$, which includes x_{n+1} as well.

Our proposed extension is quite similar to the Nystrom method [24] that has been used in spectral graph theory.

The main difference in our formulation is that we ignore the eigenvalues due to the modification described above.

7. Experimental setup and results

We evaluate the methods described above on the LFW dataset [1]. As is customary, we test the effect of the various contributions on the 10 folds of view 2 of the LFW dataset.

There are three benchmarks that are commonly used, and we provide very competitive results on all three. The most popular supervised benchmark is the "image-restricted training". This is a challenging benchmark which consists of 6,000 pairs, half of which are "same" pairs. The pairs are divided into 10 equally sized sets. The benchmark experiment is repeated 10 times, where in each repetition, one set is used for testing and nine others are used for training. The task of the tested method is to predict which of the testing pairs are matched, using only the training data (in all three benchmarks, the decision is done one pair at a time, without using information from the other testing pairs). The second supervised benchmark, constructed on top of the LFW dataset, is the "unrestricted" benchmark. In this benchmark, the persons' identities within the nine training splits are known, and the systems are allowed to use this information. For example, in this benchmark, the original WCCN method can be used directly since the training set is divided into identity-based classes. Last, the unsupervised benchmark uses the same training set. Here, however, all the training images are given as one large set of images without any pairing or label information. The evaluation task remains the same as before – distinguish between matching ("same") and non-matching ("not-same") pairs of face images.

7.1. Front-end

Our system makes no use of training data outside of the LFW dataset, except for the implicit use of outside training data through trained facial feature detectors that are used to align the images, since we use the aligned LFW-a [22] set of images. The aligned images were cropped to 150×80 pixels as suggested in [6]. In contrast to other leading contributions [5, 25, 26], we did not apply any further type of preprocessing that utilizes pose estimators or 3D modeling.

7.2. Descriptors and parameters

We evaluate 5 different descriptors: LBP, Three Patch LBP (TPLBP), OCLBP, SIFT and the Scattering descriptor. For LBP we used the same parameters that were used in [6] while for TPLBP we used the parameters reported in [12]. We used the SIFT descriptors computed by [7]. For the OCLBP descriptors, we used View 1 in order to determine the following set of configurations (see Section 3.1 for a detailed description of the OCLBP parameters):

$$S=\{(10,10,\frac{1}{2},\frac{1}{2},8,1),(14,14,\frac{1}{2},\frac{1}{2},8,2),(18,18,\frac{1}{2},\frac{1}{2},8,3)\}$$

Note that in all three scales, the horizontal and vertical overlap parameters are both set to half.

For the Scattering descriptor we used the Scattering Toolbox release from [27]. We set it to use the Gabor wavelet and the values suggested in [27]: a scattering order of 2, maximum scale of 3 and 6 different orientations.

The original descriptor dimensions are 7080, 40887, 9216, 3456 and 96520 for the LBP, OCLBP, TPLBP, SIFT and Scattering, respectively.

7.3. System parameters

We used View 1 of the dataset to determine the parameters of the system. The WPCA dimension is set to 500, the DM dimension is also set to 500 and the Gaussian kernel parameter is fixed at $\sigma = 4$. In the unrestricted and restricted benchmarks, we used LDA dimensions of 100, 100, 100, 30 and 70 for the LBP, OCLBP, TPLBP, SIFT and Scattering descriptors, respectively. As already mentioned in Section 6, we chose the threshold in the unsupervised WCCN algorithm such that the number of generated 'same' labels is 15% of all pairs.

7.4. Results

We evaluate the proposed system for each feature and its square root version under the restricted, unrestricted and unsupervised protocols. The experimental results are presented in Tables 1-6, and depict the mean classification accuracy $\hat{u}$ and standard error of the mean SE.

The unsupervised results for the individual face descriptors are depicted in Table 1. The table shows the progression from the baseline "raw" descriptors, before any learning was applied, through the use of dimensionality reduction (WPCA or DM) to the results of applying unsupervised WCCN (Section 6.1) on the dimensionality reduced descriptors. As can be seen, the suggested pipeline improves the recognition quality of all descriptors significantly, in both the dimensionality reduction step and in the unsupervised WCCN step. No clear advantage to either WPCA or DM is observed.

The results obtained by combining the facial descriptors together (excluding the original LBP descriptor) are reported in Table 4. This combination, here and throughout all fusion results in this paper, is done by a simple summation of the similarity scores using uniform weights. The table also shows, for comparison, the results of solely employing OCLBP and the best results obtained by previous works. While our face description method is considerably simpler than I-LPQ* [28], which is currently

the state of the art in this category, it outperforms it, even with the usage of a single descriptor.

Results in the supervised-restricted benchmark are reported in Table 2 for the individual features and in Table 5 for the combined features. In Table 2, we present four possibilities which differ by the dimensionality reduction algorithm used: PCA followed by LDA (PCALDA), DM followed by LDA (DMLDA), WPCA or DM. WCCN, is then applied in all four cases. As a usual trend, it seems that employing LDA in between the unsupervised dimensionality reduction (PCA or DM) and the WCCN method improves results. It is important to clarify that both LDA and WCCN were applied in a restricted manner by using only pairs information, i.e. no explicit information about the identities was used and each pair formed a mini-class of its own.

Table 5 presents the combined results of all the descriptors, excluding the original LBP descriptor (due to the use of OCLBP). The combined method ("DM+WPCA fusion") includes the four descriptors (with and without square root) and both PCA+LDA+WCCN and DM+LDA+WCCN (a total of 16 scores). It is evident that combining the DM based method together with the PCA based method improves performance over using PCA or DM separately.

In comparison to previous methods, our method outperforms the state of the art by a large margin. The only exception is the "Tom-vs-Pete" [5] method which uses an external labeled dataset, which is much bigger than the LFW dataset, and employs a much more sophisticated face alignment method. Our system considerably outperforms the accuracy of 90.57% obtained by [3] in the case of a similarity-based alignment as used by LFW-a, in spite of the fact that our method does not use the added external data.

The results for the supervised-unrestricted benchmark are depicted in Tables 3 and 6. The classical form of WCCN [10] applies directly to this setup. Two systems outperform ours in this category. The first is CMD+SLBP (aligned) which is a commercial system [29]. The second [30] has a few distinguishing characteristics, which can be further utilized to improve our results. First, a different alignment method was used. Second, features were extracted on facial landmarks. Finally, their proposed algorithm operated in a much higher-dimensional feature space, which requires more computational resources.

In all three experiments, OCLBP achieves a very competitive accuracy as a single feature. For example, as can be seen in Table 5, in the restricted case it achieves an accuracy which is much better than the current best reported accuracy obtained by [6]. The Scattering transform based description (Section 3.2), however, does not seem to improve over descriptors of lower dimensionality by a significant margin. Nevertheless, it plays a crucial role in increasing performance in fusion.

System	Accuracy
I-LPQ*, aligned [28]	86.20 ± 0.46
OCLBP	86.66 ± 0.30
WPCA fusion	88.00 ± 0.36
DM fusion	87.87 ± 0.41
DM+WPCA fusion	88.57 ± 0.37

Table 4: Comparison of classification accuracy (± standard error) for various systems operating in the unsupervised setting.

System	Accuracy
LBP + CSML, aligned [6]	85.57 ± 0.52
CSML + SVM, aligned [6]	88.00 ± 0.37
High-Throughput BIF, aligned [14]	88.13 ± 0.58
Associate-Predict [3]	90.57 ± 0.56
Tom-vs-Pete + Attribute [5]	93.30 ± 1.28
OCLBP	87.85 ± 0.69
PCA fusion	90.61 ± 0.56
DM fusion	90.26 ± 0.55
DM+PCA fusion	91.10 ± 0.59

Table 5: Comparison of classification accuracy (± standard error) for various systems operating in the restricted setting.

System	Accuracy
LBP PLDA, aligned [26]	87.33 ± 0.55
combined PLDA [26]	90.07 ± 0.51
face.com r2011b [25]	91.30 ± 0.30
CMD + SLBP, aligned [29]	92.58 ± 1.36
combined Joint Bayesian [31]	90.90 ± 1.48
high-dim LBP [30]	93.18 ± 1.07
OCLBP	88.75 ± 0.60
DM fusion	91.56 ± 0.45
PCA fusion	91.56 ± 0.54
DM+PCA fusion	92.05 ± 0.45

Table 6: Comparison of classification accuracy (± standard error) for various systems operating in the unrestricted setting.

One can also notice that the unsupervised WCCN in some of the cases achieves an accuracy which is not far away from the accuracy obtained by the original supervised WCCN. For example, for the OCLBP descriptor, WPCA + supervised WCCN achieves an accuracy of 87.2% for the restricted case while the WPCA + unsupervised WCCN pipeline achieves an accuracy of 86.7 %.

8. Conclusions

We propose an effective method that seems to be unique in that it addresses all three benchmarks in a unified manner. In all three cases, very competitive results are achieved. The method is heavily based on dimensionality reduction algorithms, both supervised and unsupervised, in order to utilize high dimensionality representations. Necessary adjustments are performed in order to adapt methods such as WCCN and DM to the requirements of face identification and of the various benchmark protocols.

From a historical perspective, our method is "reactionary". The emergence of the new face recognition benchmarks has led to the abandonment of the classical algebraic methods such as Eigenfaces and Fisherfaces. However, both PCA and LDA play important roles in our pipeline, even though these methods are not applied

directly to image intensities. WCCN, which is a major contributing component to our pipeline, was borrowed and adapted from the speaker recognition domain. However, it is closely related to other algebraic dimensionality reduction methods. In contrast to recent contributions such as CSML [6] or the Ensemble Metric Learning method [29] that are influenced by modern trends in metric learning, our method demonstrates that classical face recognition methods can still be relevant to contemporary research.

References

[1] G. B. 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, 2007.

[2] Z. Cao, Q. Yin, X. Tang and J. Sun, "Face Recognition with Learning-based Descriptor," in *Computer Vision and Pattern Recognition (CVPR)*, 2010.

[3] Q. Yin, X. Tang and J. Sun, "An Associate-Predict Model for Face Recognition," in *Computer Vision and Pattern Recognition (CVPR)*, 2011.

[4] E. Nowak and F. Jurie, "Learning visual similarity measures for comparing never seen objects," in *Computer Vision and Pattern Recognition (CVPR)*, 2007.

[5] T. Berg and P. N. Belhumeur, "Tom-vs-Pete Classifiers and Identity-Preserving Alignment for Face Verification," in *British Machine Vision Conference (BMVC)*, 2012.

[6] H. V. Nguyen and L. Bai, "Cosine Similarity Metric Learning for Face Verification," in *Asian Conference on Computer Vision (ACCV)*, 2010.

[7] M. Guillaumin, J. Verbeek and C. Schmid, "Is that you? Metric Learning Approaches for Face Identification," in *International Conference on Computer Vision (ICCV)*, 2009.

[8] N. Kumar, A. C. Berg, P. N. Belhumeur and S. K. Nayar, "Attribute and Simile Classifiers for Face Verification," in *International Conference on Computer Vision (ICCV)*, 2009.

[9] T. Ahonen, A. Hadid and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 28, p. 2037–2041, 2006.

[10] A. Hatch and A. Stolcke, "Generalized linear kernels for one-versus-all classification: application to speaker recognition," in *ICASSP*, 2006.

[11] R. Coifman and S. Lafon, "Diffusion maps," *Applied and computational harmonic analysis,* vol. 21, no. 1, pp. 5-30, 2006.

[12] L. Wolf, T. Hassner and Y. Taigman, "Descriptor based methods in the wild," in *Faces in Real-Life Images Workshop in European Conference on Computer Vision (ECCV)*, 2008.

[13] S. Mallat, "Group invariant scattering," [Online]. Available: arxiv.org/abs/1101.2286.

[14] N. Pinto and D. Cox, "Beyond Simple Features: A Large-Scale Feature Search Approach to Unconstrained Face Recognition," in *International Conference on Automatic Face and Gesture Recognition (FG)*, 2011.

[15] N. Pinto, J. J. DiCarlo and D. D. Cox, "How far can you get with a modern face recognition test set using only simple features?," in *Computer Vision and Pattern Recognition (CVPR)*, 2009.

[16] T. Ojala, M. Pietikainen and D. Harwood, "A comparative study of texture measures with classification based on feature distributions," *Pattern Recognition,* vol. 29, no. 1, pp. 51-59, 1996.

[17] M. Heikkilä, M. Pietikäinen and C. Schmid, "Description of interest regions with center-symmetric local binary patterns," in *Computer Vision, Graphics and Image Processing*, 2006.

[18] X.-M. Ren, X.-F. Wang and Y. Zhao, "An Efficient Multi-scale Overlapped Block LBP Approach for Leaf Image Recognition," *Lecture Notes in Computer Science,* vol. 7390, pp. 237-243, 2012.

[19] S. Liao, X. Zhu, Z. Lei, L. Zhang and S. Z. Li, "Learning Multi-scale Block Local Binary Patterns for Face Recognition," in *Advances in Biometrics*, 2007.

[20] J. Bruna and S. Mallat, "Invariant scattering convolution networks," 2012. [Online]. Available: http://arxiv.org/abs/1203.1513.

[21] L. Sifre and S. Mallat, "Combined scattering for rotation invariant texture analysis," in *European Symposium on Artificial Neural Networks*, 2012.

[22] Y. Taigman, L. Wolf and T. Hassner, "Multiple One-Shots for Utilizing Class Label Information," in *British Machine Vision Conference (BMVC)*, 2009.

[23] Z. N. Karam and W. M. Campbell, "Graph-embedding for speaker recognition," in *Interspeech*, 2010.

[24] C. Fowlkes, S. Belongie, F. Chung and J. Malik, "Spectral Grouping Using the Nystrom Method," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 26, no. 2, pp. 214-225, 2004.

[25] Y. Taigman and L. Wolf, "Leveraging billions of faces to overcome performance barriers in unconstrained face recognition," 2011 . [Online]. Available: http://arxiv.org/abs/1108.1122.

[26] P. Li, Y. Fu, U. Mohammed, J. H. Elder and S. J.D.Prince, "Probabilistic Models for Inference About Identity," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 34, no. 1, pp. 144-157, 2012.

[27] "Image scattering toolbox v3," [Online]. Available: http://www.di.ens.fr/data/software/.

[28] S. u. Hussain, T. Napoléon and F. Jurie, "Face Recognition Using Local Quantized Patterns," in *British Machine Vision Conference (BMVC)*, 2012.

[29] C. Huang, S. Zhu and K. Yu, "Large Scale Strongly Supervised Ensemble Metric Learning, with Applications to Face Verification and Retrieval," NEC, 2011.

[30] D. Chen, X. Cao, F. Wen and J. Sun, "Blessing of Dimensionality: High-dimensional Feature and Its Efficient Compression for Face Verification," in *Computer Vision and Pattern Recognition (CVPR)*, 2013.

[31] D. Chen, X. Cao, L. Wang, F. Wen and J. Sun, "Bayesian Face Revisited: A Joint Formulation," in *European Conference on Computer Vision (ECCV)*, 2012.