A DISSIMILARITY KERNEL WITH LOCAL FEATURES FOR ROBUST FACIAL RECOGNITION

Weilin Huang and Hujun Yin

School of Electrical and Electronic Engineering
The University of Manchester, Manchester, M60 1QD, UK.

ABSTRACT

Local binary pattern (LBP) has recently been proposed for texture analysis and local feature description and has also been applied to face recognition with promising results. However, besides the descriptors, a suitable similarity measure that can efficiently learn to distinguish facial features is also important. In this paper, a novel framework for robust face recognition is presented that considers both local and global features by using multi-resolution LBP descriptors. The framework can tolerate variations in expression, lighting condition and occlusion. A weighted distance measure is used to learn the dissimilarity between sets of LBP features. We formulate the distance function as a conditionally positive semi-definite (CPD) kernel, thus making it suitable for kernel-based algorithms such as support vector machines (SVMs) whose optimal solutions are guaranteed. We show that by defining it in a Hilbert space, the proposed CPD kernel has advantages over traditional methods computing the l_2 distances in the Euclidean space. The experiments show that the approach is efficient and significantly outperforms the current state-of-the-art methods on the publicly available AR face database.

Index Terms— Local binary pattern, local feature, dissimilarity measure, robustness, conditionally positive semi-definite kernel

1. INTRODUCTION

Face recognition process often has two main phases, representation and learning, both of which are recurrent topics in computer vision. The task of representation is to find efficient salient features from raw face images. The holistic methods play a key role in this task, such as dimensionality reduction that considers the whole image as a point in the high-dimensional input space and reduces it onto a lower dimensional output space while retaining intrinsic features. Subspace methods such as PCA [1], LDA [2] and manifold learning techniques [3] have long been considered as efficient global approaches for feature extraction. But these global approaches usually consider an image as a histogram of pixels exactly arranged by their fixed spatial locations and thus are sensitive to changes in illumination, rotation and spatial scale.

Component-based approach has recently gained attention due to its robustness to multiple facial variations. Previous work has been carried out on applying local descriptors for face representation with good performance [4]. Local binary pattern (LBP) [5] is a powerful local texture descriptor and has been recently applied to face recognition [6]. Its success is mainly due to two reasons. First, it is invariant to monotonic gray-level changes by thresholding the neighboring pixel values and therefore discarding the contrast. This also makes computation efficient. Second, face images can be considered as a composition of micropatterns that are well described by such operators. Note, though some other widely used local representations such as bag of words (BOW) [7] model with SIFT descriptor [8] have had various successes in object recognition, face recognition is somewhat a different task that requires more discriminative and detailed texture features and spatial geometric information. BOW model specializing in extracting general and orderless features discards all spatial information and therefore becomes robust to scalings and rotations. In [6], facial spatial information was enhanced by splitting image into many subregions and an LBP descriptor was applied to each subregion. Though good results were yielded and the advantages were evaluated, some excellent properties of LBP descriptor have not been fully explored. In this paper, we revisit it and highlight its gray-scale invariance property to illumination.

Defining similarity/dissimilarity measures lies in the core of machine learning. An efficient measure computed in high-dimensional feature spaces where features have more structural distribution can greatly facilitate the task of learning. An advanced approach is to construct an effective kernel function over sets of features. The kernel function selects features and measures similarity between them through the so-called *kernel* while defining an appropriate regularization term for the learning problem. Pyramid match kernel [9] maps local features to multi-resolution histograms and defines a positive semi-define kernel by computing a weighted histogram intersection. Lazebnik et al. [10] extended the pyramid kernel to spatial pyramid matching by considering global geometric information. Wolf et al. [11] constructed a CPD kernel based on the LDA classifier for one-shot similarity measure. Efficient match kernels [12] achieve more accurate similarity measures by mapping local features to a low-dimensional space and constructing set-level features.

In the proposed framework, the LBP is used for local feature extraction and further extended for considering both local and global features by using multi-resolution spatial partition. An efficient kernel function is defined for dissimilarity measure, computing the weighted distances between local features. We show that the proposed kernel is CPD and can be used in kernel-based learning machines which consider the CPD kernel as the measure of l_2 distance in a Hilbert space. Experiments were conducted on the AR database and robustness to illumination, expression and occlusion is achieved. In the next section, related work is reviewed and then the proposed framework is presented in Section 3. The experimental results and comparisons are given in Section 4, followed by conclusions in Section 5.

2. PREVIOUS WORK

2.1. Local Binary Patterns

A basic LBP descriptor assigns a label to each pixel of an image by thresholding its neighbors (g_n) into a set of binary numbers using the values of center pixel (g_c): if $g_n \geq g_c$, $s(g_n - g_c) = 1$, otherwise,

$s(g_n - g_c) = 0$, and then computing the label from these binary numbers. The image is represented as a LBP histogram containing 2^P (P is the number of neighbors) bins, each of which counts the total number of each sperate label through the image.

Ojala et al. made two extensions of LBP descriptor in [5]. First, a circle (instead of square) neighborhood is used. A set of neighbors is evenly distributed on this circle. The number of neighbors (P) and the radius (R) of the circle can be varied. The second improvement is the definition of the *uniform* patterns. A local binary is regarded as *uniform* if the number of its spatial transitions (referred as $U(pattern)$) is at most two. Otherwise the pattern is called *nonuniform*. For example, the patterns with $U(11000111) = 2$ are *uniform*, while patterns having $U(10100000) = 3$ are *nonuniform*. In computing LBP histogram, each bin counts the number of *uniform* patterns and the number of all *nonuniform* patterns is recoded by a single bin because *uniform* patterns account for high percentage (usually higher than 85%) of all patterns [5, 6]. These two extensions further improve its robustness to gray-scale changes and rotations while greatly reducing computation complexity by using only $P + 1$ bins instead of original 2^P in each LBP histogram. We follow the notations as in [5] and refer $LBP^{riu2}_{P,R}$ as the label of pixel by using circle neighborhood with *uniform* definition.

2.2. Spatial Pyramid Matching

Spatial pyramid matching [10] is an extension of the pyramid match kernel [9], which finds approximate similarities between *unordered* feature sets by computing weighted histogram intersections in the multi-resolution histogram spaces. It performs pyramid matchings by considering the rough global geometric information achieved by repeatedly subdividing an image. The image is first represented as a set of features through a local descriptor (e.g. SIFT) and each feature has a unique coordinate corresponding to its spatial location in the image. Then a visual vocabulary (including M words) is constructed by performing k-means clustering of all training features. Features are separated into M types corresponding to their nearest words in the vocabulary. Each image is repeatedly divided into 4^l subregions, $l = 0, \ldots, L$, is the level of resolution. The number of intersections (I^l) between two images, $\mathbf{X}_i$ and $\mathbf{X}_j$, at resolution l is,

$$I^l(\mathbf{X}_i, \mathbf{X}_j) = \sum_{m=1}^{M} \sum_{k=1}^{4^l} \min(X_{ik}^{lm}, X_{jk}^{lm}) \tag{1}$$

where X_{ik}^{lm} and X_{jk}^{lm} are the numbers of features (in the type of m) whose coordinates are located inside the k-th subregion.

The weight associated to the l-th resolution is set to $\frac{1}{2^{L-l}}$, inversely proportional to the size of subregion. Note that the intersections found at l-th resolution also occur at the finer $l + 1$ resolution. Therefore the number of new intersections found at l-th resolution is only $I^l - I^{l+1}$. The spatial pyramid matching is computed as

$$K^L(\mathbf{X}_i, \mathbf{X}_j) = I^L + \sum_{l=0}^{L-1} \frac{1}{2^{L-l}} (I^l - I^{l+1}) \tag{2}$$

3. PROPOSED FRAMEWORK

The proposed framework consists of multi-resolution LBP feature representation and a dissimilarity measure for kernel learning. A weighted distance function is constructed and we show that it is a CPD kernel and is suitable for kernel-based learning methods.

(a) *N1* (b) *G1* (c) *GL1* (d) *S1* (e) *SL1*

Fig. 1. Samples of LBP faces, corresponding to the original faces, *N1*, *G1*, *GL1*, *S1* and *SL1*, presented in Fig. 2.

3.1. Multi-Resolution LBP Histogram for Representation

For face representation, we consider three key elements: efficient local description, robustness to illumination variations and trade-off between local and global features. As mentioned in Section 2.1, LBP descriptor is highly efficient. For example, the descriptor used in our experiments, $LBP^{riu2}_{24,3}$, has only 25 bins in each LBP histogram, while the SIFT descriptor normally uses 128 or 256 bins. Note that these $P + 1$ types of patterns representing different types of texture features, such as lines, edges, corners and spots [5], correspond to the M feature types in spatial pyramid matching model [10]. Unlike [10] which separates the features into different types by performing complicated k-means clustering on training sets, these features can be generated directly from LBP descriptor. Examples of LBP faces are presented in Fig. 1. The prominent advantage of LBP, invariance to monotonic transformation of gray scales, can be easily seen.

As discussed in Section 1, the representation with all detailed global geometric information, such as in those holistic methods, is considered as 'overfit' description and usually leads to performance sensitive to intra-class diversities; the representation with little global geometric information (e.g. BOW [7]) is considered too 'coarse' to represent essential features such as shape of face. Here we combine local features with appreciate global geometric information by repeatedly dividing a face image into subregions and representing each subregion by a LBP histogram. The numbers of subregions at l-th resolution is 4^l and therefore a face image is represented as a set of $\sum_{l=0}^{L} 4^l$ features.

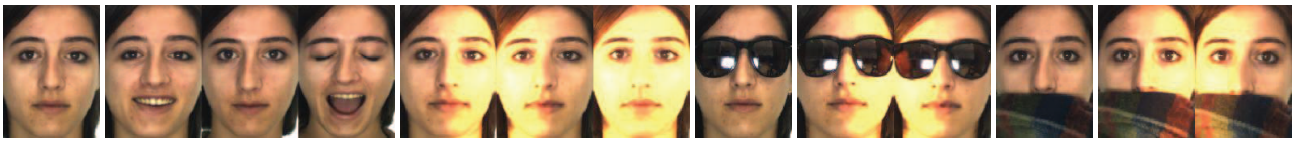
3.2. Weighted Distance between Sets of Features

Let $\mathbf{Y}_i$ and $\mathbf{Y}_j$ be two sets of LBP features for images $\mathbf{X}_i$ and $\mathbf{X}_j$ respectively. The feature histogram corresponding to l-th resolution and k-th subregion is presented as $\mathbf{Y}_{ik}^l$, and the Euclidean distance between two feature sets in this subregion is $D_k^l(\mathbf{X}_i, \mathbf{X}_j) = |\mathbf{Y}_{ik}^l - \mathbf{Y}_{jk}^l|^2$. The distance of feature sets at l-th resolution is the sum of all subregion distances,

$$D^l(\mathbf{X}_i, \mathbf{X}_j) = \sum_{k=1}^{4^l} |\mathbf{Y}_{ik}^l - \mathbf{Y}_{jk}^l|^2 \tag{3}$$

Following similar principles in [10] that distances between subregions of smaller size are more discriminative and vice versa. A weight value, which is inversely proportional to the size of subregion, is assigned to the distance at each resolution, and is computed as $\frac{1}{4^{L-l}}$, corresponding to the number of subregions, 4^l, at l-th resolution. The final distance function is

$$D^L(\mathbf{X}_i, \mathbf{X}_j) \quad = \quad D^L + \sum_{l=0}^{L-1} \frac{1}{4^{L-l}} (D^l - D^{l+1}) \tag{4}$$



(a) *N1* (b) *E1* (c) *L1* (d) *G1* (e) *GL1* (f) *S1* (g) *SL1*

Fig. 2. A subject of AR faces in session one, *N1*-Nature, *E1*-Expressions (smile, angry and scream), *L1*- Lighting (left lighting, right lighting and all sides lighting), *G1*-Sum Glasses, *GL1*-Sun Glasses & Lighting, *S1*- Scarf and *SL1* - Scarf & Lighting.

3.3. Conditionally Positive semi-Definite Dissimilarity Kernel

Though we use weighted distances to optimize discriminative power at difference resolutions, the distance in Eq. (4) is measured in Euclidean space and its application is limited to simple learning algorithms such as the nearest neighbors (NN). For using advanced methods, the distance should be defined in some specific form, e.g. positive semi-definite [13], which guarantees an optimal solution. In [14], a larger class of functions, conditionally positive semi-definite (CPD) function, is proved to be suitable for kernel methods. In the kernel theory of [14], the Hilbert space representation of a CPD kernel is the negative of a l_2 distance measure, $-|\phi(x_i) - \phi(x_j)|^2$. Therein we construct a CPD kernel for distance measure.

Definition 1 (Conditionally Positive simi-Definite Kernel). *A symmetric function $k : \chi \times \chi \longrightarrow \mathbb{R}$, for which for all $n \in \mathbb{N}$, $x_i \in \chi$, and if vectors $\boldsymbol{c} \in \mathbb{R}^n$ such that $\sum_{i=1}^{n} c_i = 0$, we have $\boldsymbol{c}^T \boldsymbol{K} \boldsymbol{c} \geq 0$, where $\boldsymbol{K} \in \mathbb{R}^{n \times n}$ is the matrix $K_{ij} = k(x_i, x_j)$, and is called a* conditionally positive definite (CPD) kernel.

From Definition 1, we can deduce that l_2 distance function is not CPD, because $\sum_i c_i = 0$, this implies $\sum_{i,j} c_i c_j |x_i - x_j|^2 = \sum_i c_i \sum_j c_j |x_j|^2 + \sum_j c_j \sum_i c_i |x_i|^2 - 2\sum_{i,j} c_i c_j \langle x_i, x_j \rangle = -2\sum_{i,j} c_i c_j \langle x_i, x_j \rangle = -2|\sum_i c_i x_i|^2 \leq 0$. However, negative l_2 distance kernel, $k(x_i, x_j) = -|x_i - x_j|^2$, is a CPD kernel.

Proposition 1 (Fractional Powers and Logs of CPD Kernels [13]). *If $\psi : \chi \times \chi \longrightarrow (-\infty, 0]$ is CPD, then so are $-(-\psi)^\alpha (0 < \alpha < 1)$ and $-ln(1 - \psi)$.*

Following Definition 1, we know that (*i*) a sum of CPD kernels is CPD , (*ii*) any constant $C \in \mathbb{R}$ is a CPD kernel, and (*iii*) a multiplication of CPD kernels with any non-negative constant $C \geq 0$ is a CPD kernel. According to Proposition 1, we obtain a general form of CPD kernel for distance functions,

$$k(x_i, x_j) = C - |x_i - x_y|^\beta \qquad 0 < \beta < 2, C \in \mathbb{R} \tag{5}$$

Combining Eqs. (3)-(5), we have the final CPD kernel for the dissimilarity measure,

$$\begin{aligned} K^L(\mathbf{X}_i, \mathbf{X}_j) &= C - \frac{1}{4^L} |\mathbf{Y}_{i1}^0 - \mathbf{Y}_{j1}^0|^\beta \\ &\quad - \sum_{l=1}^{L} \sum_{k=1}^{4^l} \frac{3}{4^{L-l+1}} |\mathbf{Y}_{ik}^l - \mathbf{Y}_{jk}^l|^\beta \end{aligned} \tag{6}$$

4. EXPERIMENTAL RESULTS AND COMPARISONS

Two experiments on the AR face database were conducted for evaluating the proposed framework. In the first, only one standard (nature) face of each subject was used for training and the remaining for testing. We compared the performance with that of holistic methods to show the advantages of multi-resolution local methods and the capability of the LBP descriptor under lighting variations. We also benchmarked the dissimilarity measure in SVMs against traditional kernels. In the second experiment, the proposed method was compared with two of the state-of-the-arts methods on the same dataset, same set-up and same traing/testing schemes.

Table 1. Classification results for AR face database

Test	Correction rates(%)					
	EigF	FishF	LBP	LBP-P	LBP-G	LBP-WD
	NN			*SVM*		
E	82.83	73.33	80.83	84.67	82.83	**95.83**
L	13.00	62.33	86.33	91.33	87.17	**95.33**
G	42.00	48.50	84.00	89.50	82.50	**99.50**
GL	33.25	22.75	76.00	78.75	71.75	**93.75**
S	6.50	17.50	41.50	51.50	44.50	**88.50**
SL	13.00	12.00	30.75	33.25	26.50	**69.00**

The AR database consists of over 3000 color images of 126 subjects, each having 26 facial images taken in two different sessions separated by two weeks. Each session has 13 images with multiple variations in expression, illumination and occlusion (sun glasses and scarf). A subset of cropped faces of 50 male and 50 female subjects [15] was used, the same subset used by [16, 17]. Examples of a subject from one session are shown in Fig. 2.

4.1. Experiment One

In this experiment, multi-resolution local LBP descriptors ($LBP_{24,3}^{riu2}$) are used for feature representation, compared with Eigenfaces (EigF) and Fisherfaces (FishF) methods. LBP histograms were normalized by dividing the cardinality of $LBP_{24,3}^{riu2}$ at each resolution. In learning and classification, NN, SVM with polynomial (LBP-P) and Gaussian(LBP-G) kernels, and the proposed distance kernel (LBP-WD) were evaluated. LBP-WD was computed by Eq. (6) with $\beta = 1$, $C = 1$ and $L = 5$. Two independent implementations were conducted, both trained on nature faces (*N*) and tested on remaining images from the same session, grouped by expressions (*E*), lighting (*L*), sun glasses (*G*), sun glasses with lighting (*GL*), scarf (*S*) and scarf with lighting (*SL*). All 100 subjects were used and results presented in Table 1 are the mean values of two implementations.

The results show that, for feature representation, though holistic methods have good performance in expression, they are very sensitive to changes in lighting and occlusion. It shows that detailed geometric features of holistic representation are intolerant to these changes. The other reason is that only one image per subject is used for training. As expected, the original local LBP descriptor yields better results on lighting and sun glasses separately and good results

Table 2. Comparisons with [16, 17]

Testing groups	**Training/testing scheme**		**Correction rates**(%)	
Experiment [16]	Training	Testing	Results of [16]	Our results
expre. & light.	7×100 images: *N1, E1, L1*	7×100 images: *N2, E2, L2*	94.70	**97.86**
sun glasses	8×100 images: *N1, N2, E1, E2*	2×100 images: *G1, G2*	97.50	**100.00**
scarf	8×100 images: *N1, N2, E1, E2*	2×100 images: *S1, S2*	93.50	**96.50**
Experiment [17]	Training	Testing	Results of [17]	Our results
scream	6×50(100) images: *N1, N2, E1(1,2), E2(1,2)*	2×50(100) images: *E1(3), E2(3)*	87.00	**98.50**
sun glasses	6×50(100) images: *N1, N2, E1(1,2), E2(1,2)*	2×50(100) images: *G1, G2*	84.00	**100.00**
scarf	6×50(100) images: *N1, N2, E1(1,2), E2(1,2)*	2×50(100) images: *S1, S2*	93.00	**96.50**

on case affected by both. Although it outperforms holistic methods, it still performs poorly on large occlusions. For comparison of dissimilarity kernels, SVMs with LBP-WD kernels considerably outperform SVMs with the traditional kernels, as shown in Table 1. The proposed kernel with local LBP descriptor improves the performance significantly in all groups. It appears very robust to variations on expression, lighting and small occlusion. Though the rate falls to 69% when faces are seriously affected by both large occlusion and lighting, it is still far better than other methods compared.

4.2. Experiment Two

We compared our approach on efficiency with two of most recent methods in the literature [16] and [17], both of which are based on holistic representation. We compared all the available results of them on the dataset with same training/testing schemes detailed in Table 2. Note that all 100 subjects were used in [16] and our experiments, but only 50 subjects were used in [17]. The results show that our approach significantly outperform both methods in all testing groups in accuracy. Neither of the experiments in [16] and [17] tested the groups of faces affected by multiple factors such as GL or SL, in which our method achieved reasonable to excellent results as shown in Table 1. Holistic methods such as Eigenface and Fisherface perform poorly when only few samples are available for training, and many methods require that sufficient training samples are available either for representation [16] or for reconstruction [17]. The success of holistic methods heavily depend on training samples covering sufficient cardinality and variation. However, in practical applications, obtaining 'sufficient' samples can be costly or even impossible. A method based on only few training data is highly desirable.

5. CONCLUSIONS

A novel framework for robust face recognition is proposed by considering both feature representation and leaning metric. The property of local LBP descriptor is further exploited for dealing with illuminance changes. A multi-resolution partition is used to effectively combine both local and global features for robust representation and has achieved significant improvements over traditional holistic feature representation. A CPD kernel function is defined for dissimilarity measure by computing weighted distances between sets of multi-resolution LBP features. It can be used in kernel learning algorithms by defining l_2 distance in the feature space. The experimental results show that the CPD kernel with the enhanced LBP descriptor markedly improves the performance over traditional kernels. The proposed method is also compared with two recent influential approaches and the results show its advantages in accuracy, efficiency and applicability with few training samples.

6. REFERENCES

[1] M. Turk and A. Pentland, "Eigenfaces for recognition," *J. of Cognitive Neuroscience*, vol. 3, pp. 71–86, 1991.

[2] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. on PAMI*, vol. 19, pp. 711–720, 1997.

[3] Y. Goldberg, A. Zakai, D. Kushnir, and Y. Ritov, "Manifold learning: The price of normalization," *J. of Machine Learning Research*, vol. 9, pp. 1909–1939, 2008.

[4] W. Zhao, R. Chellappa, P. J. Phillips, and A. A. Rosenfeld, "Face recognition: A literature survey," *ACM Computing Surveys*, vol. 35, pp. 399–458, 2003.

[5] T. Ojala, M. Pietikäinen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. on PAMI*, vol. 24, pp. 971–987, 2002.

[6] A. Ahonen, A. Hadid, and M. Pietikäinen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. on PAMI*, vol. 28, pp. 2037–2041, 2006.

[7] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *ICCV*, 2003, pp. 1470–1477.

[8] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, pp. 91–110, 2004.

[9] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," in *ICCV*, 2005, pp. 1458 – 1465.

[10] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006, pp. 2169 – 2178.

[11] L. Wolf, T. Hassner, and Y. Taigman, "The one-shot similarity kernel," in *ICCV*, 2009.

[12] L. Bo and C. Sminchisescu, "Efficient match kernel between sets of features for visual recognition," in *NIPS 22*, 2009, pp. 135–143.

[13] C. Berg, J. P. R. Christensen, and P. Ressel, *Harmonic Analysis on Semigroups*, Springer-Verlag, New York, 1984.

[14] B. Schölkopf, "The kernel trick for distances," in *NIPS 13*, 2000, pp. 301–307.

[15] A. M. Martinez and A. C. Kak, "Pca versus lda," *IEEE Trans. on PAMI*, vol. 23, pp. 228–233, 2001.

[16] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. on PAMI*, vol. 31, pp. 210–227, 2009.

[17] S. Fidler, D. Skočaj, and A. Leonardis, "Combining reconstructive and discriminative subspace methods for robust classification and regression by subsampling," *IEEE Trans. on PAMI*, vol. 28, pp. 337–350, 2006.