Pattern Recognition Letters
journal homepage: www.elsevier.com

Local Directional Texture Pattern Image Descriptor

Adín Ramírez Rivera[a,**], Jorge Rojas Castillo[b], Oksam Chae[b]

[a] *Escuela de Informática y Telecomunicaciones, Universidad Diego Portales, Ejército 441, Santiago, 8320000, Chile*
[b] *Department of Computer Engineering, Kyung Hee University, 1 Seocheon-dong, Giheung-gu, Yongin-si, Gyeonggi-do 446-701, South Korea*

ABSTRACT

Deriving an effective image representation is a critical step for a successful automatic image recognition application. In this paper, we propose a new feature descriptor named Local Directional Texture Pattern (LDTP) that is versatile, as it allows us to distinguish person's expressions, and different landscapes scenes. In detail, we compute the LDTP feature, at each pixel, by extracting the principal directions of the local neighborhood, and coding the intensity differences on these directions. Consequently, we represent each image as a distribution of LDTP codes. The mixture of structural and contrast information makes our descriptor robust against illumination changes and noise. We also use Principal Component Analysis to reduce the dimension of the multilevel feature set, and test the results on this new descriptor as well.

1. Introduction

Nowadays several applications need to recognize something through visual cues, such as faces, expressions, objects, or scenes. These applications need robust descriptors that discriminate between classes, but that are general enough to incorporate variations within the same class. However, most of the existing algorithms have a narrow application spectrum, *i.e.*, they focus on some specific task. For example, previous methods (Jabid et al., 2010; Ojala et al., 1996) designed for face analysis, cannot be used to recognize scenes, as these methods have been tuned for fine texture representation. Similarly, descriptors (Lazebnik et al., 2006; Wu and Rehg, 2011) tuned for scene recognition under-perform on other tasks. Thereby, in this paper, we design and test a robust descriptor capable of modeling fine textures as well as coarse ones.

A wide range of algorithms have been proposed to describe micro-patterns. The most common ones are the appearance-based methods (Shan et al., 2009) that use image filters, either on the whole-image, to create holistic features, or some specific region, to create local features, to extract the appearance change in the image. The performance of the appearance-based methods is excellent in constrained environment but their performance degrade in environmental variation (Pantic and Rothkrantz, 2000). In the literature, there are many methods for the holistic class, such as Eigenfaces and Fisherfaces. Although these methods have been studied widely, local descriptors have gained attention because of their robustness to illumination and pose variations. The Local Binary Pattern (LBP) (Ojala et al., 1996) is by far the most popular one, and has been successfully applied to several problem domains (Shan et al., 2009; Zhao et al., 2012; Zhou et al., 2008). Despite LBP's robustness to monotonic illumination, it is sensitive to non-monotonic illumination variation, and shows poor performance in presence of random noise (Zhou et al., 2008). Tan and Triggs (2007) proposed an improvement of LBP by introducing a ternary pattern (LTP) which uses a threshold to stabilize the micro-patterns. A directional pattern (LDP) (Jabid et al., 2010) has been proposed to overcome the limitations of LBP. However, it suffers in noisy conditions, is sensitive to rotations, and cannot detect different transitions in the intensity regions. Similarly, many other methods appeared that extract information and encoded it in a similar way like LBP, such as infrared (Xie and Liu, 2011), near infrared (Zhang et al., 2010), and phase information (Chan et al., 2009; Zhang et al., 2007). Nevertheless, all these methods inherit the sensitivity problem, *i.e.*, the feature being coded into the bit-string is prone to change due to noise or other variations. Thereby, the directional-number-based methods (Ramirez Rivera et al., 2012a,b,c; Rojas Castillo et al., 2012) appeared as a solution to the common bit-string

**Corresponding author: Tel.: +56-02-26768134;
e-mail: adin.ramirez@mail.udp.cl (Adín Ramírez Rivera)

representation, as these methods use an explicit coding scheme in which the prominent directions are embedded into the code. However, all these methods still encode only one type of information, *e.g.*, intensity or direction, which limits their description capabilities.

Therefore, in this paper, we propose a novel feature descriptor named Local Directional Texture Pattern (LDTP) which exploits the advantages of both directional and intensity information in the image. The combination of both features outperforms the singled-feature counterparts, *e.g.*, LBP, LTP, LDP, among others. The proposed method identifies the principal directions from the local neighborhood, and then extracts the prominent relative intensity information. Following, LDTP characterizes the neighborhood by mixing these two features in a single code. On the contrary, previous methods rely on one type of information and use a sensitive coding strategy. In detail, LDTP encodes the structural information in a local neighborhood by analyzing its directional information and the difference between intensity values of the first and second maximum edge's responses. This mechanism is consistent against noise, since the edge response is more stable than intensity, and the use of relative intensity values makes our method more robust against illumination changes and other similar conditions. Moreover, given that we encode only the prominent information of the neighborhood, our method is more robust to changes in comparison to other methods, as we dispose insignificant details that may vary between instances of the image. Consequently, we convey more reliable information of the local texture, rather than coding all the information that may be misleading, not important, or affected by noise. Furthermore, we evaluate the performance of the proposed LDTP feature with a machine learning method, Support Vector Machine, on five different databases for expression recognition and three different databases for scene recognition to demonstrate its robustness and versatility.

2. Local Directional Texture Pattern

LBP feature labels each pixel by thresholding a set of sparse points of its circular neighborhood, and encodes that information in a string of bits. Similarly, LDP encodes the principal directions of each pixel's neighborhood into an eight bit string. However, these methods only mark which neighbor has the analyzed characteristic on or off. Furthermore, this *ad hoc* construction overlooks the prominent information in the neighborhood, as all the information in the neighborhood (regardless of its usefulness) is poured into the code. In contrast, we create a code from the principal directions of the local neighborhood, similar to the directional numbers (Ramirez Rivera et al., 2012a,b,c; Rojas Castillo et al., 2012). However, the later has the problem of using only the structural information of the neighborhood. Therefore, we propose to extract the contrast information from the principal directions to enhance the description of our code. In other words, we code the principal direction and the intensity difference of the two principal directions into one number. This approach allows us to encode the prominent texture information of the neighborhood that is revealed by its principal directions.

To compute LDTP, we calculate the principal directional numbers of the neighborhood using the Kirsch compass masks (Kirsch, 1970)—in eight different directions. We define our directional number as

$$P^1_{\text{dir}} = \arg\max_i \{\mathbb{I}_i \mid 0 \le i \le 7\}, \tag{1}$$

where P^1_{dir} is the principal directional number, $\mathbb{I}_i$ is the absolute response of the convolution of the image, I, with the ith Kirsch compass mask, M_i, defined by

$$\mathbb{I}_i = |I * M_i|. \tag{2}$$

Thus, we compute the absolute value of the eight Kirsch mask's responses, $\{M_0, \ldots, M_7\}$, applied to a particular pixel. More precisely, we take the two greatest responses, P^1_{dir} and P^2_{dir}. Therefore, the second directional number, P^2_{dir}, is computed in the same way, with the difference that we take the second maximum response in Eq. 1 instead. These directions signal the principal axis of the local texture.

In each of the two principal directions, we compute the intensity difference of the opposed pixels in the neighborhood. That is

$$d^{(x,y)}_n = I(x_{P^n_{\text{dir}}+}, y_{P^n_{\text{dir}}+}) - I(x_{P^n_{\text{dir}}-}, y_{P^n_{\text{dir}}-}), \tag{3}$$

where d_n is the nth difference for the pixel (x, y) in the nth principal direction, $I(x_{P^n_{\text{dir}}+}, y_{P^n_{\text{dir}}+})$ corresponds to the intensity value of the pixel $(x_{P^n_{\text{dir}}+}, y_{P^n_{\text{dir}}+})$, which is the next pixel in the given principal direction, and $I(x_{P^n_{\text{dir}}-}, y_{P^n_{\text{dir}}-})$ is the intensity value of the pixel $(x_{P^n_{\text{dir}}-}, y_{P^n_{\text{dir}}-})$, which is the previous pixel in the given principal direction. In other words, the next and previous pixel positions defined by each direction are

$$x_{P^n_{\text{dir}}\pm} = \begin{cases} x \pm 1 & \text{if } P^n_{\text{dir}} \in \{0, 1, 7\}, \\ x & \text{if } P^n_{\text{dir}} \in \{2, 6\}, \\ x \mp 1 & \text{if } P^n_{\text{dir}} \in \{3, 4, 5\}, \end{cases} \tag{4}$$

$$y_{P^n_{\text{dir}}\pm} = \begin{cases} y \pm 1 & \text{if } P^n_{\text{dir}} \in \{1, 2, 3\}, \\ y & \text{if } P^n_{\text{dir}} \in \{0, 4\}, \\ y \mp 1 & \text{if } P^n_{\text{dir}} \in \{5, 6, 7\}. \end{cases} \tag{5}$$

This local difference, is equivalent to the local threshold that LBP does. Unlike the LBP binary encoding, we encode the difference using three levels (negative, equal, and positive), which creates a more distinctive code for the neighborhood. Then each difference is encoded as

$$D_f(d) = \begin{cases} 0, & \text{if } -\varepsilon \le d \le \varepsilon \\ 1, & \text{if } d < -\varepsilon \\ 2, & \text{if } d > \varepsilon, \end{cases} \tag{6}$$

where D_f is the encoded intensity difference, d is the actual intensity difference, ε is a threshold value (in our experiments we use $\varepsilon = 15$).

Consequently, the code is created by concatenating the binary form of the principal direction, and the two differences. This concatenation can be represented by the following operation

$$\text{LDTP}(x, y) = 16P^{1(x,y)}_{\text{dir}} + 4D_f(d^{(x,y)}_1) + D_f(d^{(x,y)}_2), \tag{7}$$

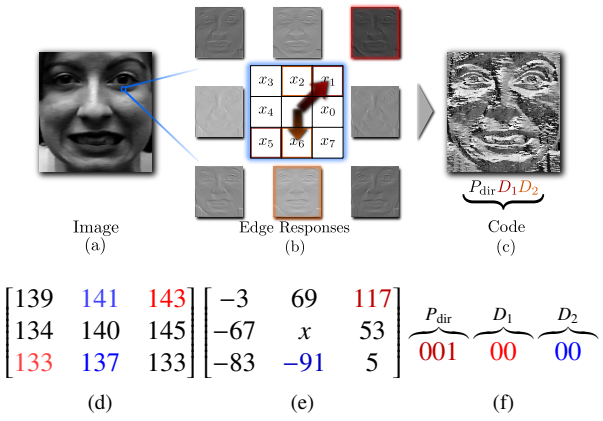

Fig. 1. Example of the LDTP code computation (details are in the text). (a) Original image. (b) Edge responses after applying Kirsch masks. (c) Coded image. (d) Sample neighborhood (intensity values). (e) Edge responses of the sample neighborhood shown in (d). (f) Code of the neighborhood shown in (d).

where LDTP(x, y) is the code for the pixel (x, y), $P_{\text{dir}}^{1(x,y)}$ is the principal directional number (from 0 to 7) of the neighborhood of the pixel (x, y), and $D_f(d_1^{(x,y)})$ and $D_f(d_2^{(x,y)})$ are the first and second coded differences of the neighborhood of the pixel (x, y), respectively. The length of the code is $72 = 8 \times 3 \times 3$, as the possible values for the directional number is 8, and 3 for each difference.

For example, consider the neighborhood shown in Fig. 1(d), first we compute the Kirsch mask responses in the neighborhood—we show them in their respective orientation in Fig. 1(e). The principal, M_1, and the secondary, M_6, directions are shown in red and blue, respectively. Then, we compute the intensity difference of the corresponding pixel intensities in these directions [as shown by the colored pairs in Fig. 1(d)]. In this case, the differences are: $d_1 = 143 - 133 = 10$, and $d_2 = 137 - 141 = -4$, which are transformed with Eq. 6 into $D_f(d_1) = 0$, and $D_f(d_2) = 0$, assuming a threshold $\varepsilon = 15$. Finally, we create the LDTP code by concatenating the binary form of the principal direction index, and the two differences as shown in Fig. 1(f).

If we opt to include more information into our code by embedding the two principal directional numbers [like previous methods (Ramirez Rivera et al., 2012a)], then we will increase its discrimination power too much as well as its length. Thus, we will diminish its recognition capabilities, because most of the textures will be coded differently. Consequently, we decided to maintain a compact code by using the principal directional information and the contrast information of the two principal ones.

3. Image descriptor using LDTP

We represent the images through a set of histograms of LDTP features. Consequently, we generate the LDTP coded image by using Eq. 7. Each code contains micro patterns of the image, which represent certain information of each neighborhood. However, the histogram loses the spatial information of the coded image. Hence, we divide the image into several regions, $\{R_0, \ldots, R_N\}$, and compute a histogram H_i for each region R_i, where each bin corresponds to a different pattern value—note that in the LDTP feature we have 72 different possible code values. Furthermore, to construct the final descriptor we concatenate the histograms of each region, R_i, into a single feature vector through

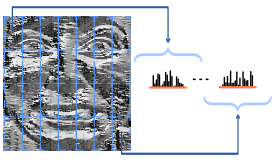

Fig. 2. Face representation using combined LDTP histogram.

$$\mathbb{H} = \bigmathbin{\Big\|}_{i=0}^{N} H_i, \tag{8}$$

where $\mathbb{H}$ is the final descriptor, H_i is the histogram of the LDTP codes for region R_i, and $\|$ is the concatenation operation. An example of the different regions and the histogram concatenation is shown in Fig. 2.

4. Experimental Setup

In order to evaluate the performance of the proposed encoding algorithm, we performed experiments in two different areas: facial expression and scene recognition. Despite these two being different domain fields, they have a common approach for recognition. Both use descriptors and classifiers to identify its objects of interest. However, the main difference is that the facial expression recognition needs a descriptor that describes the micro patterns, while scene recognition needs robustness to changes and variations. Therefore, we evaluated these two scenarios to demonstrate the versatility and robustness of our proposed LDTP, which can describe micro-patterns while maintaining the robustness in presence of challenging variations. In the following we explain the setup for the different experiments.

4.1. Facial expression recognition

We perform person-independent facial expression recognition, by using Support Vector Machines (SVM) to classify the coded images. SVM (Cortes and Vapnik, 1995) is a supervised machine learning technique that implicitly maps the data into a higher dimensional feature space. Consequently, it finds a linear hyperplane, with a maximal margin, to separate the data in different classes in this higher dimensional space.

Given that SVM makes binary decisions, multi-class classification can be achieved by adopting the one-against-one or one-against-all techniques. In our work, we opt for one-against-one technique, which constructs $k(k-1)/2$ classifiers, that are trained with data from two classes (Hsu and Lin, 2002). We perform a grid-search on the hyper-parameters in a 10-fold cross-validation scheme for parameter selection, as suggested by Hsu et al. (2003). The parameter setting producing the best cross-validation accuracy was picked.

To evaluate the methods, we test the facial expression recognition problem on the Cohn-Kanade Facial Expression (CK)

database (Kanade et al., 2000). For 6-class prototypical expression recognition, the three most expressive image frames were taken from each sequence that resulted into 1224 expression images. In order to build the neutral expression set, the first frame (neutral expression) from all 408 sequences was selected to make the 7-class expression data set (1632 images). Furthermore, we used the extended Cohn-Kanade database (CK+) (Lucey et al., 2010), which includes sequences for seven basic expressions. In our experiments, we selected the most expressive image frame (the last frame is the apex of the sequence in this database) from 325 sequences from 118 subjects from the database for evaluation. These sequences are the ones with correct labels, all the other sequences are not correctly labeled or have a missing labels from the database. Additionally, we used the Japanese Female Facial Expression (JAFFE) database (Lyons et al., 1998), which contains only 213 images of female facial expression expressed by ten subjects. Moreover, we tested the expression recognition problem on the MMI face database (Valstar and Pantic, 2010). In our experiments we used the Part II of the database, which comprises 238 clips of 28 subjects (sessions 1767 to 2004) where all expressions (anger, disgust, fear, happiness, sadness, and surprise) were recorded twice. We also used the CMU-PIE database (Sim et al., 2003), which includes 41368 face images of 68 people captured under 13 poses, 43 illuminations conditions, and with four different expressions: neutral, smile, blinking, and talk. For our experiments, we tested two expressions: smile and neutral, as blinking and talking requires temporal information, which is out of the scope of this publication. Moreover, we used the poses that are near frontal (camera 27) with horizontal (cameras 05 and 29) and vertical rotation (cameras 07 and 09).

For our experiments, we cropped all the images to 110×150 pixels, based on the ground truth positions of the eyes and mouth, and partitioned the images into 5×5 regions. Since LDTP detects the principal components of the textures, no further alignment of facial features was performed in our data, and since it is robust against illumination changes, no attempts were made for pre-processing in order to remove such changes. Note that these actions were taken to demonstrate the robustness of the proposed method against the corresponding changes. Moreover, we achieve person-independent classification by dividing the databases into several partitions and by ensuring that one person's expressions are not distributed into two partitions. In our experiments, we randomly divide the images into ten partitions, and we performed a leave-one-out cross-validation, which is equivalent to a 10-fold cross validation.

4.2. Scene recognition

Similarly to the face recognition problem, we represent the scene as a set of histogram of LDTP features. Consequently, we generate the LDTP coded scene using Eq. 7. In contrast to the face analysis, we incorporate more spatial information, to improve classification performance, by adopting spatial pyramid matching scheme, as Lazebnik et al. (2006) did. They captured more spatial information by concatenating each micro-block of the pyramid's levels. Consequently, it conveys more important features of the scenes in different resolutions, as Fig. 3 shows.

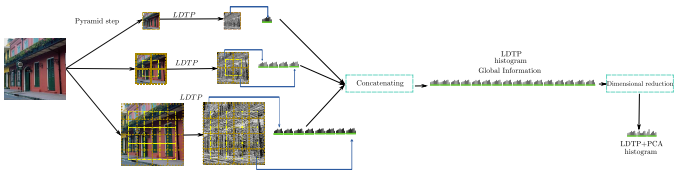

Fig. 3. Steps for creating the histogram of the scene by using a spatial pyramid.

We normalized the pictures to 256×256 pixels, and exploit the spatial information by creating a three level pyramid. The idea behind the pyramid is to capture changes in scale that may appear in different scenes. Thus, the first level has one block, the second level has four non-overlapping blocks plus one overlapping block, and the third level has 16 non-overlapping blocks plus nine overlapping blocks. The use of the overlapping blocks is to capture sets of features that may lie in the intersection of blocks. Moreover, the images are re-sized between different levels so that all the blocks contain the same number of pixels. In total, we have 31 blocks for each image; and we create a global representation of the image by concatenating the descriptor of each block. The final descriptor is computed using Eq. 8.

For scene recognition we used the datasets provided by Fei-Fei and Perona (2005), Lazebnik et al. (2006), and Oliva and Torralba (2001) to evaluate the performance of the algorithms. The first database contains eight scenes categories provided by Oliva and Torralba (2001): mountain (274 images), coast (360 images), highway (260 images), street (292 images), insidecity (308 images), forest (328 images), opencountry (410 images), and tallbuilding (356 images), where the size of each image is 256×256. The second one contains fifteen scene categories, and is an extension of the first database by adding seven new scenes categories: PARoffice (215 images), living-room (289 images), bedroom (216 images), kitchen (210 images), CALsuburb (241 images), store (315 images), and industrial (311 images). The first five classes are provided by Fei-Fei and Perona (2005), and the other two are collected by Lazebnik et al. (2006). Additionally, we used the MIT Indoor Scene dataset comprised of 15620 images over 67 indoor scenes assembled by Quattoni and Torralba (2009). We follow their experimental setting by using 80 images for training and 20 for testing.

5. Experimental Results

We evaluated the performance of the proposed method with the images from several databases (as explained in Section 4). To further improve the detection rate, we used PCA (Joliffe, 1986) to discover low dimensional feature of LDTP, and we also present its results here as LDTP+PCA.

5.1. (Extended) Cohn-Kanade results

The recognition rates of the proposed methods (LDTP and LDTP with PCA) in comparison with other methods—Local Binary Pattern (LBP) (Shan et al., 2009), Local Directional Pattern (LDP) (Jabid et al., 2010), Gabor features (Bartlett et al., 2003), and Local Ternary Pattern (Tan and Triggs, 2007)—are

Table 1. Comparison against others methods, in CK and JAFFE databases.

Method	CK		JAFFE	
	6 class (%)	7 class (%)	6 class (%)	7 class (%)
LBP	92.6 ± 2.9	88.9 ± 3.5	86.7 ± 4.1	80.7 ± 5.5
LDP	98.5 ± 1.4	94.3 ± 3.9	85.8 ± 1.1	85.9 ± 1.8
Gabor	89.8 ± 3.1	86.8 ± 3.1	85.1 ± 5.0	79.7 ± 4.2
LTP	99.3 ± 0.2	92.5 ± 2.5	80.3 ± 1.0	77.9 ± 1.0
LDTP	99.4 ± 1.1	95.1 ± 3.1	90.2 ± 1.0	88.7 ± 0.5
LDTP+PCA	**99.7 ± 0.9**	**95.7 ± 2.9**	**92.4 ± 1.2**	**89.2 ± 0.8**

Table 2. Confusion matrix of 6-class facial expression recognition using SVM (RBF) with LDTP in the CK database.

(%)	Anger	Disgust	Fear	Joy	Sadness	Surprise
Anger	99.5					0.5
Disgust		100				
Fear			100			
Joy				100		
Sadness	2.1				97.9	
Surprise						100

shown in Table 1. The LDTP codes perform better in the 6- and 7-class problem on the CK database. To obtain a better picture of the recognition accuracy of individual expression types, we present the confusion matrices for 6- and 7-class expression recognition using the CK database for the best LDTP codes in Tables 2 and 3. These results show that the 6-class recognition problem can be solved with high accuracy (as we have a miss detection of 2% between the sadness and anger expressions). However, as we include the neutral expression in the 7-class recognition problem, the accuracy of recognizing five expressions decreases, as the descriptor cannot differentiate between expression displays that are too mild. This effect is more evident as the surprise expression is not confused, as it involves the rise and opening of the eyes, which differentiates it greatly from other expressions.

In the CK+ dataset, we compared our descriptor against several geometric-base methods. Canonical appearance features (CAPP) and similarity-normalized shape (SPTS) proposed by Lucey et al. (2010) with the CK+ dataset. Moreover, Chew et al. (2011) proposed a constrained local model (CLM) based method. Also, Jeni et al. (2011) proposed a CLM method by using shape related information only (CLM-SRI). Furthermore, we compared against a method based on emotion avatar image (EAI) (Yang and Bhanu, 2012) that leverages the out of plane rotation. Additionally, we test against the LTP (Tan and Triggs,

Table 3. Confusion matrix of 7-class facial expression recognition using SVM (RBF) with LDTP in the CK database.

(%)	Anger	Disgust	Fear	Joy	Sadness	Surprise	Neutral
Anger	87.5				1.3	0.5	10.7
Disgust		97					3
Fear	0.5		97				2.5
Joy			0.4	98.8			0.8
Sadness					95.2		4.8
Surprise						100	
Neutral	2.7	0.5	0.5		1.3		95

Table 4. Recognition accuracy (%) for expressions on the (a) CK+, (b) MMI, and (c) CMU-PIE.

(a)

Method	CK+
SPTS	50.4
CAPP	66.7
SPTS+CAPP	83.3
CLM	74.4
EAI	82.6
LTP	78.7
LDTP	81.4
LDTP+PCA	**84.5**

(b)

Method	MMI
LBP	86.9
CPL	49.4
CSPL	73.5
AFL	47.7
ADL	47.8
LTP	89.5
LDTP	90.5
LDTP+PCA	**93.7**

(c)

Method	CMU
LBP	85.1
LBP_w	90.3
LTP	88.8
LDP	88.4
LPQ	90.9
LDTP	90.9
LDTP+PCA	**92.9**

Table 5. Confusion matrix of 7-class recognition using SVM (RBF), in the CK+ database.

(%)	Anger	Contempt	Disgust	Fear	Happy	Sadness	Surprise
Anger	67.5		7.5	2.5	2.5	17.5	2.5
Contempt		93.75					6.25
Disgust	14.29		82.54	1.59			1.59
Fear	12.5	4.17		70.83		12.5	
Happy		1.45			98.55		
Sadness	21.43	3.57		14.29		60.71	
Surprise			4.6	2.3		1.15	91.95

2007) method—notice that in our experiments we did not use any preprocessing which degrades the recognition accuracy of the LTP method as it relies solely on intensity. Table 4(a) shows that our methods outperform all the other, even thought they are geometric based, which use a more complex representation of the face. Yet, our proposed LDTP outperforms them with a simple representation. Jeni et al. (2011) used a temporal normalization step which yields an accuracy of 96%. However, for a fair comparison against all the other methods we leave this score outside of the table, and used the result that do not use the temporal normalization.

Additionally, we present the confusion matrix of our LDTP descriptor in Table 5. The worst confusion occurs for anger, fear, and sadness expressions. These expressions have small changes in them that difficult the representation from a single frame. For example, the sadness expression can be seen as anger as the mouth and eyebrows present similar shape and position. To improve the accuracy temporal information may be used to input more cues to identify the expression being performed. Nevertheless, the expressions with high structural display, such as contempt, happiness, and surprise were detected with high accuracy.

Overall, the high accuracy of LDTP is due to the use of prominent information for the coding process. That is, LDTP extracts the information from the principal axis of the texture of each local neighborhood. In contrast, other methods use all the information in the neighborhood, which may be sub-optimal or influenced by noise.

5.2. Different-resolution facial expression recognition

In real world environments, such as smart meeting and visual surveillance, the resolution of the images is not high. Thus, we investigate these scenarios by changing the resolutions of the

Table 6. Recognition rate (%) in low-resolution (CK database) images.

Resolution	Methods			
	LBP	LDP	Gabor	**LDTP+PCA**
110 × 150	92.6 ± 2.9	98.5 ± 1.4	89.8 ± 3.1	**99.7 ± 0.9**
55 × 75	89.9 ± 3.1	95.5 ± 1.6	89.2 ± 3.0	**98.8 ± 3.1**
36 × 48	87.3 ± 3.4	93.1 ± 2.2	86.4 ± 3.3	**98.5 ± 1.4**
27 × 37	84.3 ± 4.1	90.6 ± 2.7	83.0 ± 4.3	**95.1 ± 1.3**
18 × 24	79.6 ± 4.1	89.8 ± 2.3	78.2 ± 4.5	**94.8 ± 1.3**
14 × 19	76.9 ± 5.0	89.1 ± 3.1	75.1 ± 5.1	**94.2 ± 2.7**

Cohn-Kanade database images. The images were down sampled from the original images to 110×150, 55×75, 36×48, 27×37, 18×24, and 14×19. We present the experiment results with six different resolutions on Table 6. As seen in the table, the proposed code is more robust than other methods under resolution changes. LDTP has an improvement of 3.7% in average over the second best method in each resolution. Again, this high performance comes from the extraction and use of the prominent information in the neighborhood, and our novel coding scheme.

5.3. JAFFE results

We compared against the same methods used in the CK experiment for the JAFFE experiment. We observed that the recognition accuracy in JAFFE database, shown in Table 1, is relatively lower than the CK database. One of the main reasons behind this accuracy is that some expressions in the JAFFE database are very similar with each other. Thus, depending on whether these expression images are used for training or testing, the recognition result is influenced. Nevertheless, our method still outperforms other methods due to the extraction of the principal components to form each local code.

5.4. MMI results

To evaluate the proposed methods on the MMI database, we compared them against two recent studies: a boosted LBP (Shan et al., 2009) and several patch-based approaches based on the former method (Zhongy et al., 2012). Zhongy et al. (2012) proposed two methods Common Patches (CPL) and Common and Specific Patches (CSPL) with LBP to produce a more localized descriptor. Moreover, they use Adaboost (ADL) to learn certain patches in the face, and code them using LBP; also they use all available patches (AFL) to create the descriptor and recognize the expressions. We also compared against the Local Ternary Pattern (LTP) (Tan and Triggs, 2007). Table 4(b) shows that the proposed method outperforms previous methods by 4.2% to the second best.

For a better comprehension of the performance of our approach on the MMI database, Table 7 shows the confusion matrix of the LDTP descriptor. We note that, from all the expressions, the fear expression gets confused with surprise, and disgust. This confusion is due to the similarity among the expressions, as some people only rise their eyebrows when surprised, while others open their mouth, which may lead to some confusion. Hence, to improve this detection, temporal information may be incorporated.

Table 7. Confusion matrix of 6-class recognition using SVM (RBF) in MMI database.

(%)	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	96.49				3.51	
Disgust		89.13	4.35	2.17	4.35	
Fear		4.55	84.09			11.36
Joy		4.76	1.59	93.65		
Sadness	1.85		1.85		96.3	
Surprise			9.72			90.28

5.5. CMU-PIE results

In the CMU-PIE dataset Sim et al. (2003) we focused on a two class classification problem: smile and neutral. To evaluate this experiment we compared our proposed method against several local descriptors: LBP (Ahonen et al., 2006), LBP_w (Xie and Liu, 2011), LTP (Tan and Triggs, 2007), LDP (Jabid et al., 2010), and LQP (Chan et al., 2009).

We show on table 4(c) the results of all the different descriptors in this dataset. We found that the variation in the head poses influences the result. Moreover, our method achieves the same performance as the second best method, while applying PCA improves the result on the recognition by 2%.

5.6. Scene recognition results

Similarly to the face analysis, we used PCA to reduce the number of bins in the histogram. After reducing the data, we used 5-fold cross validation and an RBF kernel to classify scene images. Finally, we report the average value.

We evaluated our methods against several methods that we describe in the following. Oliva and Torralba (2001) proposed the Gist descriptor to represent the similar spatial structures present in each scene category. Gist computes the spectral information in an image through Discrete Fourier Transform, and the spectral signals are then compressed by the Karhunen-Loeve Transform. The CENsus TRansform hISTogram (CENTRIST) (Wu and Rehg, 2011) is a holistic representation and has strong generalizability for category recognition. CENTRIST mainly encodes the structural properties within an image and suppresses detailed textural information. Additionally, we compared our methods against Lazebnik et al. (2006) method: the spatial pyramid matching (SPM). In one variation (SPM-1), they used 16 channels weak features, where their recognition rate is 66.8%; and in the other experiment (SPM-2), they increased their recognition rate by 14.6% because they incorporated the SIFT descriptor using 400 visual words. Liu and Shah (2007) used SIFT, as previous method (Lazebnik et al., 2006) did, with the difference that they used 400 intermediate concepts to convey more reliable information (SPM-C). Similarly, Bosch et al. (2008) used SIFT and 1200 pLSA topics to incorporate more information (SP-pLSA).

Tables 8(a) and 8(b) show the comparison on the 8- and 15-class scene data sets, in which our proposed method (LDTP+PCA) outperforms other methods. Despite the improvement being small, our method is more versatile as it can work on fine textures as well as large image representations. Additionally, we did not preprocessed the images to extract

Table 8. Comparison of the recognition accuracy of the methods for (a) 8 and (b) 15 scene categories.

(a)

Method	Recognition (%)
Gist	82.60 ± 0.86
CENTRIST	85.65 ± 0.73
LDTP	81.18 ± 0.78
LDTP+PCA	**85.81 ± 0.67**

(b)

Method	Recognition (%)
SPM-1	66.80 ± 0.60
SPM-2	81.40 ± 0.50
SPM-C	83.25 ± N/A
SP-pLSA	83.70 ± N/A
Gist	73.28 ± 0.67
CENTRIST	83.88 ± 0.76
LDTP	73.12 ± 0.69
LDTP+PCA	**83.94 ± 0.85**

Table 9. Recognition accuracy of the methods for the MIT indoor scenes.

Method	Recognition (%)
ROI+Gist	25.0
OB	37.6
CENTRIST	36.9
LPC	39.6
LDTP	**38.1**
LDTP+PCA	35.7

parts of the images to train our descriptor, as we used the images as a whole which speeds up the process.

Additionally, we tested our proposed method in the MIT indoor scene dataset against the ROI annotation and Gist method proposed by Quattoni and Torralba (2009), the Object Bank (OB) proposed by Li et al. (2010), and the Local Pairwise Codebook (LPC) proposed by Morioka and Satoh (2010). Table 9 shows the results of different methods in this dataset. Our proposed method has a similar accuracy to other state of the art methods. However, it is not the best method in this challenging dataset—*c.f*. LPC (Morioka and Satoh, 2010). Nevertheless, our method is still better than other similar methods such as CENTRIST. We noted that the use of PCA in this experimental setting did not boost the result, due to the amount of different scenes in the training and testing partitions. Due to the amount of diversity in the scenes the principal components extracted by PCA did not better represent the data. Moreover, we tested a combined framework using CENTRIST and our proposed method. In CENTRIST authors used an LBP descriptor, and we combined it by concatenating ours and LBP descriptor in the CENTRIST framework. Thus, in that environment we found an accuracy of 41.5% in the MIT indoor scene dataset.

6. Conclusion

In this paper, we proposed a novel local image descriptor based on Local Directional Texture Pattern (LDTP) code, that can work in a wide variety of scenarios. LDTP extracts the texture information from the principal axis in each neighborhood, and thus encodes the prominent characteristics of such neighborhood. The main distinction with existing methods is that instead of trying to accommodate all available information, which sometimes may introduce errors into the code, LDTP includes only the principal information of the micro-pattern. Consequently, our proposed descriptor can accommodate a large variety of problems. In this paper, we explored its use for facial expression recognition and scene recognition, and showed that LDTP can achieve a higher recognition accuracy over existing methods in the tested databases.

Acknowledgments

This work was supported in part by CONICYT, grant funded by the Chilean Government, under FONDECYT Iniciación No. 11130098.

References

Ahonen, T., Hadid, A., Pietikäinen, M., 2006. Face description with local binary patterns: Application to face recognition. IEEE Trans. Pattern Anal. Mach. Intell. 28, 2037–2041. doi:10.1109/TPAMI.2006.244.

Bartlett, M.S., Littlewort, G., Fasel, I., Movellan, J.R., 2003. Real time face detection and facial expression recognition: Development and applications to human computer interaction., in: Computer Vision and Pattern Recognition Workshop, 2003. CVPRW '03. Conference on, p. 53. doi:10.1109/CVPRW.2003.10057.

Bosch, A., Zisserman, A., Munoz, X., 2008. Scene classification using a hybrid generative/discriminative approach. IEEE Trans. Pattern Anal. Mach. Intell. 30, 712–727.

Chan, C.H., Kittler, J., Poh, N., Ahonen, T., Pietikäinen, M., 2009. (Multiscale) Local phase quantisation histogram discriminant analysis with score normalisation for robust face recognition, in: Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on, pp. 633–640. doi:10.1109/ICCVW.2009.5457642.

Chew, S., Lucey, P., Lucey, S., Saragih, J., Cohn, J., Sridharan, S., 2011. Person-independent facial expression detection using constrained local models, in: Automatic Face Gesture Recognition and Workshops (FG 2011), 2011 IEEE International Conference on, pp. 915 –920. doi:10.1109/FG.2011.5771373.

Cortes, C., Vapnik, V., 1995. Support-vector networks. Machine Learning 20, 273–297. URL: http://www.springerlink.com/index/10.1007/BF00994018.

Fei-Fei, L., Perona, P., 2005. A bayesian hierarchical model for learning natural scene categories, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, pp. 524 – 531 vol. 2. doi:10.1109/CVPR.2005.16.

Hsu, C.W., Chang, C.C., Lin, C.J., 2003. A practical guide to support vector classification. Taipei, Taiwan. URL: http://www.csie.ntu.edu.tw/\~{}cjlin/papers/guide/guide.pdf.

Hsu, C.W., Lin, C.J., 2002. A comparison of methods for multiclass support vector machines. IEEE Trans. Neural Netw. 13, 415–425. doi:10.1109/72.991427.

Jabid, T., Kabir, M.H., Chae, O., 2010. Facial expression recognition using local directional pattern (LDP), in: Image Processing (ICIP), 2010 17th IEEE International Conference on, pp. 1605–1608. doi:10.1109/ICIP.2010.5652374.

Jeni, L., Takacs, D., Lorincz, A., 2011. High quality facial expression recognition in video streams using shape related information only, in: Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on, pp. 2168 –2174. doi:10.1109/ICCVW.2011.6130516.

Joliffe, I.T., 1986. Principal component analysis. Springer-Verlag .

Kanade, T., Cohn, J.F., Tian, Y.L., 2000. Comprehensive database for facial expression analysis, in: Automatic Face and Gesture Recognition, 2000. Proceedings. Fourth IEEE International Conference on, pp. 46–53. doi:10.1109/AFGR.2000.840611.

Kirsch, R.A., 1970. Computer determination of the constituent structure of biological images. Computers & Biomedical Research , 315–328.

Lazebnik, S., Schmid, C., Ponce, J., 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, pp. 2169–2178. doi:10.1109/CVPR.2006.68.

Li, L.J., Su, H., Xing, E.P., Li, F.F., 2010. Object bank: A high-level image representation for scene classification & semantic feature sparsification, in: NIPS, pp. 1378–1386.

Liu, J., Shah, M., 2007. Scene modeling using co-clustering. Computer Vision, IEEE International Conference on 0, 1–7. doi:http://doi.ieeecomputersociety.org/10.1109/ICCV.2007.4408866.

Lucey, P., Cohn, J., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I., 2010. The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression, in: Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, pp. 94–101. doi:10.1109/CVPRW.2010.5543262.

Lyons, M., Akamatsu, S., Kamachi, M., Gyoba, J., 1998. Coding facial expressions with gabor wavelets, in: Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on, pp. 200 –205. doi:10.1109/AFGR.1998.670949.

Morioka, N., Satoh, S., 2010. Building compact local pairwise codebook with joint feature space clustering, in: Daniilidis, K., Maragos, P., Paragios, N. (Eds.), Computer Vision - ECCV 2010. Springer Berlin Heidelberg. volume 6311 of *Lecture Notes in Computer Science*, pp. 692–705. URL: http://dx.doi.org/10.1007/978-3-642-15549-9_50, doi:10.1007/978-3-642-15549-9_50.

Ojala, T., Pietikäinen, M., Harwood, D., 1996. A comparative study of texture measures with classification based on feature distributions. Pattern Recognition 29, 51–59.

Oliva, A., Torralba, A., 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. International Journal of Computer Vision 42, 145–175.

Pantic, M., Rothkrantz, L., 2000. Automatic analysis of facial expressions: the state of the art. IEEE Trans. Pattern Anal. Mach. Intell. 22, 1424–1445. doi:10.1109/34.895976.

Quattoni, A., Torralba, A., 2009. Recognizing indoor scenes, in: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), pp. 413–420.

Ramirez Rivera, A., Rojas Castillo, J., Chae, O., 2012a. Local directional number pattern for face analysis: Face and expression recognition. IEEE Trans. Image Process. .

Ramirez Rivera, A., Rojas Castillo, J., Chae, O., 2012b. Local gaussian directional pattern for face recognition, in: International Conference on Pattern Recognition (ICPR), pp. 1000–1003.

Ramirez Rivera, A., Rojas Castillo, J., Chae, O., 2012c. Recognition of face expressions using local principal texture pattern, in: Image Processing, 2012. ICIP 2012. IEEE International Conference on.

Rojas Castillo, J., Ramirez Rivera, A., Chae, O., 2012. Facial expression recognition based on local sign directional pattern, in: Image Processing, 2012. ICIP 2012. IEEE International Conference on.

Shan, C.F., Gong, S.G., McOwan, P.W., 2009. Facial expression recognition based on local binary patterns: A comprehensive study. Image and Vision Computing 27, 803–816. URL: http://www.sciencedirect.com/science/article/pii/S0262885608001844, doi:10.1016/j.imavis.2008.08.005.

Sim, T., Baker, S., Bsat, M., 2003. The CMU pose, illumination, and expression database. IEEE Trans. Pattern Anal. Mach. Intell. 25, 1615–1618. doi:10.1109/TPAMI.2003.1251154.

Tan, X., Triggs, B., 2007. Enhanced local texture feature sets for face recognition under difficult lighting conditions, in: Zhou, S., Zhao, W., Tang, X., Gong, S. (Eds.), Analysis and Modeling of Faces and Gestures. Springer Berlin Heidelberg. volume 4778 of *Lecture Notes in Computer Science*, pp. 168–182. URL: http://dx.doi.org/10.1007/978-3-540-75690-3_13, doi:10.1007/978-3-540-75690-3_13.

Valstar, M.F., Pantic, M., 2010. Induced disgust, happiness and surprise: an addition to the mmi facial expression database, in: Proceedings of Int'l Conf. Language Resources and Evaluation, Workshop on EMOTION, Malta. pp. 65–70. URL: http://ibug.doc.ic.ac.uk/media/uploads/documents/EMOTION-2010-ValstarPantic-CAMERA.pdf.

Wu, J., Rehg, J.M., 2011. CENTRIST: A visual descriptor for scene categorization. IEEE Trans. Pattern Anal. Mach. Intell. 33, 1489–1501. doi:http://doi.ieeecomputersociety.org/10.1109/TPAMI.2010.224.

Xie, Z., Liu, G., 2011. Weighted local binary pattern infrared face recognition based on weber's law, in: Image and Graphics (ICIG), 2011 Sixth International Conference on, pp. 429–433. doi:10.1109/ICIG.2011.51.

Yang, S., Bhanu, B., 2012. Understanding discrete facial expressions in video using an emotion avatar image. IEEE Trans. Syst., Man, Cybern. B PP, 1–13. doi:10.1109/TSMCB.2012.2192269.

Zhang, B., Shan, S., Chen, X., Gao, W., 2007. Histogram of gabor phase patterns (HGPP): A novel object representation approach for face recognition. IEEE Trans. Image Process. 16, 57–68. doi:10.1109/TIP.2006.884956.

Zhang, B., Zhang, L., Zhang, D., Shen, L., 2010. Directional binary code with application to polyu near-infrared face database. Pattern Recognition Letters 31, 2337–2344. doi:10.1016/j.patrec.2010.07.006.

Zhao, G., Ahonen, T., Matas, J., Pietikäinen, M., 2012. Rotation-invariant image and video description with local binary pattern features. IEEE Trans. Image Process. 21, 1465–1477. doi:10.1109/TIP.2011.2175739.

Zhongy, L., Liuz, Q., Yangy, P., Liuy, B., Huangx, J., Metaxasy, D.N., 2012. Learning active facial patches for expression analysis, in: Computer Vision and Pattern Recognition, 2012. Proceedings. IEEE Conference on.

Zhou, H., W., R., Wang, C., 2008. A novel extended local-binary-pattern operator for texture analysis. Information Sciences 178, 4314–4325. URL: http://www.sciencedirect.com/science/article/pii/S0020025508002715, doi:10.1016/j.ins.2008.07.015.