

# Spatiotemporal Directional Number Transitional Graph for Dynamic Texture Recognition

Adín Ramírez Rivera, *Member, IEEE*, and Oksam Chae, *Member, IEEE*

**Abstract**—Spatiotemporal image descriptors are gaining attention in the computer vision community for better representation of dynamic textures. In this paper, we introduce a dynamic-micro-texture descriptor, *i.e.*, Spatiotemporal Directional Number Transitional Graph (DNG), which describes both the spatial structure and motion of each local neighborhood by capturing the direction of natural flow in the temporal domain. We use the structure of the local neighborhood, given by its principal directions, and compute the transition of such directions between frames. Moreover, we present the statistics of the direction transitions in a transitional graph, which acts as a signature for a given spatiotemporal region in the dynamic texture. Furthermore, we create a sequence descriptor by dividing the spatiotemporal volume into several regions, computing a transitional graph for each of them, and represent the sequence as a set of graphs. Our results validate the robustness of the proposed descriptor in different scenarios for expression recognition and dynamic texture analysis.

**Index Terms**—Directional number, dynamic texture, facial expression, spatiotemporal descriptors, transitional graph.

## 1 INTRODUCTION

**D**YNAMIC textures are present in the real world, in the form of waves, fire, smoke, clouds, etc. Furthermore, consistent spatiotemporal motion, such as facial expressions, orderly pedestrian crowds, and vehicular traffic, can be seen as a generalization of dynamic textures and can be similarly represented. Moreover, the ability to discriminate dynamic patterns based on visual cues affects several applications, such as human-computer interaction, biometrics, psychology, surveillance, and video retrieval and indexing [1], [2].

The analysis of spatiotemporal textures involves several challenges (*e.g.*, combination of motion and appearance features, insensitivity to illumination, and computational simplicity). To address these problems, we propose a new spatiotemporal descriptor called a Directional Number Transitional Graph (DNG) that creates a signature for the dynamic patterns by aggregating the spatiotemporal directional changes of the dynamic-micro-textures. As the principal directional-indexes provide information in the local neighborhoods, the temporal changes in these directions (between consecutive time steps) may identify the dynamic texture that generated such changes. We can encode the transition of directions between frames as a set of transitions of directional numbers. The accumulated transitions represent unique characteristics of a pattern. Consequently, we model

transitions in a graph that acts as a signature for the texture by aggregating the changes from one directional number to another. For more complex dynamic patterns, such as facial expressions, we propose a sequence descriptor using the set of graphs extracted from a spatiotemporal grid placed over the sequence. Thus, DNG is a robust and general descriptor that models dynamic textures as well as complex dynamic patterns, such as facial expressions.

### 1.1 Related work

In contrast to static textures, dynamic textures vary not only in the spatial distribution of texture elements, but also with regard to organization and dynamics over time. Some existing methods that model the dynamics of image sequences are optical flow, spatiotemporal geometry, spatiotemporal filtering, spatiotemporal transforms, and model-based methods [2].

Optical-flow-based methods [3], [4] are popular due to their efficiency and characterization of dynamic patterns. Some methods [4] combine normal flow and periodicity features to characterize the magnitude, directionality, and periodicity of motion. Others [3] use spatiotemporal multi-resolution histograms based on velocity and acceleration fields calculated by a structure tensor method. In contrast to the goal of our proposed descriptor, recognition is highly tuned to a particular spatial appearance. Furthermore, optical flow and its normal flow component assume brightness constancy and local smoothness, which are generally difficult to justify. However, these methods model motion only, ignoring texture and appearance.

Several dynamic texture-recognition methods represent the global spatiotemporal variations in texture as a linear dynamic system. Most previous efforts, however,

- A. Ramírez Rivera is with Escuela de Informática y Telecomunicaciones, Universidad Diego Portales, Santiago, Chile (e-mail: adin.ramirez@mail.udp.cl).
- O. Chae is with the Department of Computer Engineering, Kyung Hee University, South Korea (e-mail: oschae@khu.ac.kr).
- This work was supported by a grant funded by CONICYT, under FONDECYT Iniciación No. 11130098, and by the Technological Innovation R&D Program (S2176380) funded by the Small and Medium Business Administration (SMBA, Korea).

have limited their experimentation to cases where the pattern samples are collected from the same viewpoint. Therefore, much of the performance is dependent on the spatial appearance captured by these models rather than the underlying dynamics [5], [6]. To address this issue, more complex models have emerged [7]. Despite these efforts, these methods suffer when the sequences present non-overlapping views, *i.e.*, they are not shift-invariant.

Nearly all of the research on dynamic texture recognition has considered textures to be homogeneous, *i.e.*, the spatial locations of the image regions are not considered. In this light, Zhao *et al.* [8] proposed the use of local binary patterns (LBP) in three orthogonal planes (TOP) to overcome the computational complexity of the volumetric descriptor. Similar methods that used other features, like phase quantization [9], or Fourier transformations [10] have been proposed. A recent work [11] uses nine planes to analyze the volume characteristics. Norouznezhad *et al.*'s work [11] relies on the cross-sections of several planes to compute the Histogram of Oriented Gradients. However, these algorithms use complex methods to encode the volumetric data [8], and, unlike our proposed method, depend on cross-sections of the temporal and spatial domains to reduce such complexity at the cost of decoupling the motion and appearance dynamics. Consequently, we avoid the use of cross-sections by considering the relations of entire neighborhoods and include that information into our graph-based descriptor.

Similarly, previous research on facial expression recognition focused on static images [12]. For appearance-based features, researchers used static micro-pattern descriptors to extract the appearance of the faces, using various techniques including Gabor-wavelets [12], LBP [13], local directional patterns [14], and directional numbers [15]–[18] among others. Recent research [19] focused on the analysis of dynamic image sequences to extract facial expressions, as temporal information may enhance recognition accuracy. For instance, several geometry-based dynamic feature methods have been proposed [20], [21]. In contrast, volume features [8], [22]–[24] appeared as the extension of appearance-based features in a dynamic environment, in which the image sequence is modeled as a dynamic texture. However, most of these methods produced their codes from cross-sections of the spatial and temporal data, and then recombined the data through histogram concatenation in later coding stages (*i.e.*, TOP-based methods).

## 1.2 Contribution

Current methods capture motion and spatial information separately, and combine them through concatenation of histograms into a final descriptor. Therefore, our main contribution is a new descriptor that models the dynamic information in image sequences by jointly representing the structure and motion of each micro-pattern. That is, our method captures the transition frequency between

features, thus simultaneously modeling the spatial and temporal information. Our contributions are summarized as follows. First, we propose a spatiotemporal version of the directional numbers [15]–[18] to code the salient directions of each local neighborhood, which avoids the common LBP-like bit string marking codes that are sensitive to changes in the data being coded. Second, we experiment with several sets of masks [25], [26] to extract the principal directions and analyze the results. Further, we propose a new 3D mask to extract the spatiotemporal directional responses in nine symmetry planes of a volume, which outperforms existing masks in our experiments. Finally, we use a weighted directed graph to model the changes in the directional numbers for a given region.

## 2 DIRECTIONAL NUMBER GRAPH

A directional number [15]–[18] of a local neighborhood is the index of the direction containing important information. However, for spatiotemporal images, the direction does not necessarily lie in the spatial domain. Thus, we can extend the notion of direction to the spatiotemporal domain (see Section 2.2). In previous works, directional numbers were computed using the edge response from masks in the spatial domain. In this paper, we propose a new approach to compute directional numbers and subsequently embed spatiotemporal information. Specifically, we are proposing a 3D compass mask that extracts the spatiotemporal edge response over nine planes of symmetry for the given volume.

Current methods compute a histogram on each cell of a grid on the cross-sections of the data as a descriptor. However, this histogram represents spatiotemporal relations only at the grid level, ignoring lower level relations. In contrast, we propose a graph-based approach as a new sequence descriptor, which models the changes in the directional numbers (at low levels) over time in a given region of the volume. Figure 1 shows an abstraction of the proposed method.

### 2.1 Compass masks

In order to extract the directional number responses, we use different compass masks in 2D and 3D, which give the principal directions of the neighborhood in the spatial and spatiotemporal domain, respectively. For the former, we chose a Kirsch compass mask [25] to compute the spatial directional response of a given neighborhood in eight different directions. However, any 2D mask could be used. Ideally, this mask allows the analysis of the spatial directional changes over time by creating a transitional graph over the spatial directional numbers.

On the other hand, we can extend the directional analysis to the spatiotemporal domain by using a 3D mask. We propose to use a cubic mask ( $3 \times 3 \times 3$ ) that computes the plane response of the 3D neighborhood. We consider the nine planes of symmetry of the cube to extract the primitive motion patterns represented by

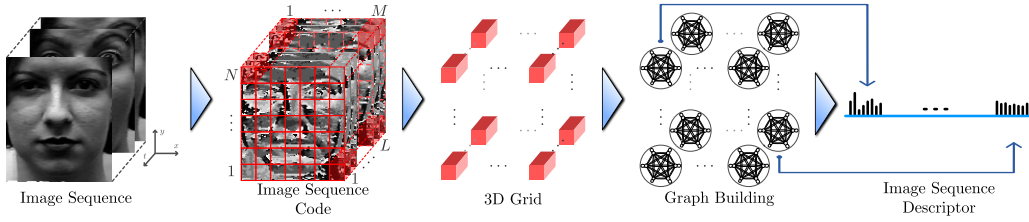


Fig. 1: We extract the spatiotemporal directional numbers for each frame, and divide the sequence into a 3D grid. For each defined region, we extract a DNG. Finally, the graphs are transformed into a one-dimensional vector to create the image sequence descriptor.

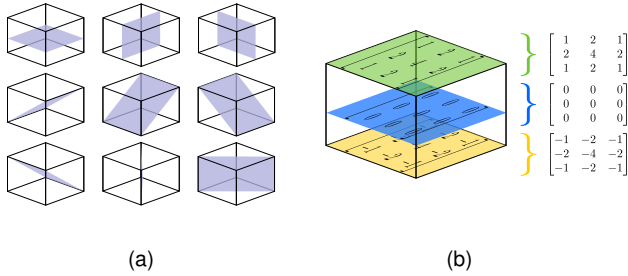


Fig. 2: (a) A 3D compass mask gives nine spatiotemporal directional responses corresponding to each of the symmetry planes of a cube. (b) Approximation of the mask that gives the XT-plane response.

the spatiotemporal directional numbers given by the compass mask shown in Fig. 2(a). Thus, we create the compass mask with a Gaussian-like weight on the direction of interest over the differencing planes, given by

$$M_1^{z=1} = \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix}, \quad (1)$$

$$M_1^{z=2} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad (2)$$

$$M_1^{z=3} = \begin{bmatrix} -1 & -2 & -1 \\ -2 & -4 & -2 \\ -1 & -2 & -1 \end{bmatrix}, \quad (3)$$

where  $M_1^{z=1,2,3}$  denotes each  $z$  plane of the cube matrix  $M_1$  that results in nine masks. Figure 2(b) shows the coefficients of the XY-plane mask ( $M_1$ ), and the other masks are the result of rotating  $M_1$  by  $45^\circ$  about each axis.

Thereby, each 3D mask gives the difference over the sides of its central plane and emphasizes the central-normal direction of the plane. We also experimented using the 3D Sobel mask proposed by Jetto *et al.* [26] in order to evaluate the efficacy of our mask. In contrast to our proposed mask, the 3D Sobel only emphasizes the central point, which limits its performance, as discussed in Section 3.

## 2.2 Transitional graph

Given a volume (spatiotemporal image)  $I$  and a (spatial or spatiotemporal) compass mask  $M = \{M_m \mid 1 \leq m \leq n\}$ , which can compute  $n$  different edge directions, we define a directional number as the index  $m$  of the

mask  $M_m$  that has a significant edge response. Thus, the directional number provides prominent information for a given neighborhood. A salient feature of our method is that, in the spatiotemporal domain, a local maximum gradient on the space-time domain indicates the relevant motion and spatial features. Thus, we create a map of salient information by extracting the principal directional number for the volume through

$$i_{x,y,t} = \arg \max_m \{\mathbb{I}_m(x, y, t) \mid 1 \leq m \leq n\}, \quad (4)$$

where  $i_{x,y,t}$  is the principal directional number for the voxel  $(x, y, t)$ , and  $\mathbb{I}_m$  is the result of convolving the volume  $I$  with the  $m$ th mask  $M_m$  through

$$\mathbb{I}_m = I * M_m. \quad (5)$$

Note that the convolution may be performed frame by frame in the case of a 2D compass mask or in the volume using a 3D mask (see Section 2.1 for details regarding the masks).

We analyze the salient changes in the volume over time by tracking the changes in the directional numbers of a given pixel over time. That is, we examine the transitions of the directional numbers in the voxels of the volume. Thus, we define a transitional graph  $G = (V, E, w)$  for the volume  $I$ , comprising the set of vertices  $V = \{v_m \mid 1 \leq m \leq n\}$  equal to the possible directional numbers defined by the compass masks, the directed edges  $E = \{(v_j, v_k) \mid \forall v_j, v_k \in V\}$ , and the weight function  $w : E \rightarrow \mathbb{R}^+$  that assigns a real number to each edge  $e \in E$ . Consequently, the transitional graph is a weighted and directed graph over the possible directional numbers defined by the compass mask. Thus, we learn the behavior of salient features by studying their changes over time and compiling statistics of these changes in the graph's weights. Moreover, the frequency of the changes acts as a signature of the volume's dynamic patterns, as different dynamic textures will produce different salient features with distinct changes.

Therefore, we determine the weights of the graph  $G$ , for all  $v_j, v_k \in V$ , using

$$w(v_j, v_k) = \sum_{(x,y,t) \in I} \delta_{j,k}(i_{x,y,t}, i_{x,y,t+1}), \quad (6)$$

where  $w(v_j, v_k)$  is the weight from vertex  $v_j$  to  $v_k$ , and

$$\delta_{j,k}(j', k') = \begin{cases} 1, & \text{if } j = j' \wedge k = k' \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

determines whether the given directional numbers  $j'$  and  $k'$  change according to the given values  $j$  and  $k$ . That is, the weight measures the frequency of the voxels that change their directional number  $i_{x,y,t}$  to  $i_{x,y,t+1}$ . In practice, the accumulation of the changes from one directional number to another can be performed in the adjacency matrix that represents the graph. Thus, each row and column is accessed using the index of the directional number of the represented vertex, and the graph can be efficiently constructed while scanning the volume. Moreover, we normalize the resulting adjacency matrix to accommodate variations in the region sizes using

$$w'(v_j, v_k) = \frac{w(v_j, v_k)}{N_r M_r L_r}, \quad (8)$$

where  $w'$  is the normalized weight,  $w$  is the original weight, and  $N_r$ ,  $M_r$ , and  $L_r$  are the dimensions of the region  $r$  being analyzed. In general, we obtain an  $8 \times 8$  adjacency matrix of the corresponding graph using the Kirsch mask, and another  $9 \times 9$  adjacency matrix for both 3D masks.

### 2.3 Sequence Descriptor

To create the sequence descriptor, we first divide the volume into  $R = NML$  different blocks  $\{B_1, \dots, B_R\}$  (as shown in Fig. 1). Secondly, we compute the transitional graph,  $G_r$ , for each block  $B_r$ . Moreover, we transform the graph's adjacency matrix,  $A_r$ , into a vector,  $\hat{A}_r$ , by concatenating the rows of the matrix, without any loss of generality. The sequence descriptor  $D$  is the concatenation of all the adjacency matrices of the partitioned volume, such that

$$D = \left\| \left\| \hat{A}_r, \right. \right\|_{r=1}^R, \quad (9)$$

where  $\|$  is the concatenation operator. We can accommodate more information by aggregating the descriptors of lower-order directional numbers. That is, we include, for example, the second principal directional number into the descriptor by recomputing Eq. 9 using the second maximum in the computation of the directional number, Eq. 4, instead. Thus, if we denote the sequence descriptor of the  $o$ th directional number (first, second, etc.) by  $D_o$ , the overall descriptor is

$$\mathbb{D} = \left\| \left\| D_o, \right. \right\|_{o=1}^O, \quad (10)$$

where  $O$  is the maximum order considered to produce the descriptor. In our experiments, we used up to a second-order descriptor, *i.e.*, we used the first and second directional numbers ( $O = 2$ ), because the use of two principal axes per neighborhood better characterizes several patterns. Consequently, the analysis and collection of the changes in the principal axis over time reveals an underlying signature of the dynamic pattern. The final descriptor represents the transition frequencies between

TABLE 1: Accuracy (%) of DNG using different SVM kernels (Inter.,  $\chi^2$ , and RBF), and a histogram descriptor (Hist.) using SVM (with an RBF kernel) and nearest neighbor (NN) classifiers on the texture datasets.

Mask	Classifier	8-Class	9-Class	SIR	DynTex++	Avg.
Kirsch	Inter.	98.0	97.6	81.3	91.0	91.9
	$\chi^2$	97.9	98.7	82.3	91.1	92.5
	RBF	97.7	99.1	81.3	92.2	<b>92.6</b>
	Hist. (SVM)	94.3	93.0	67.8	67.9	80.8
	Hist. (NN)	96.6	97.9	71.0	68.6	83.5
Sobel	Inter.	99.1	98.6	89.5	92.5	94.9
	$\chi^2$	99.1	98.0	89.5	92.7	94.8
	RBF	98.8	99.2	89.0	92.9	<b>95.0</b>
	Hist. (SVM)	96.7	96.7	84.3	80.2	89.5
	Hist. (NN)	95.8	97.0	84.5	79.3	89.2
9 Planes	Inter.	99.0	98.7	90.8	93.0	<b>95.4</b>
	$\chi^2$	99.8	98.8	89.5	93.1	95.3
	RBF	99.4	99.6	89.0	93.8	<b>95.4</b>
	Hist. (SVM)	97.1	96.8	85.3	81.0	90.1
	Hist. (NN)	94.8	96.2	84.5	79.7	88.8

TABLE 2: Dynamic texture classification accuracy (%) for the (a) UCLA and (b) DynTex++ databases.

Method	(a)			(b)		
	8-Class	9-Class	SIR	Method	(%)	
PEGASOS [6]	99.0	95.6	N/A	PEGASOS [6]	63.7	
BoS [7]	80.0	N/A	N/A	LBP-TOP [11]	71.2	
DFS [27]	99.0	97.5	73.8	DFS [27]	89.9	
HOG-NSP [11]	98.7	98.1	78.2	HOG-NSP [11]	90.1	
<b>DNG<sub>K</sub></b>	SVM	97.7	99.1	81.3	<b>DNG<sub>K</sub></b> SVM	92.2
	NN	97.7	96.9	81.3	<b>DNG<sub>K</sub></b> NN	89.8
<b>DNG<sub>S</sub></b>	SVM	98.8	99.2	<b>89.0</b>	<b>DNG<sub>S</sub></b> SVM	92.9
	NN	96.3	97.6	83.0	<b>DNG<sub>S</sub></b> NN	89.8
<b>DNG<sub>P</sub></b>	SVM	<b>99.4</b>	<b>99.6</b>	<b>89.0</b>	<b>DNG<sub>P</sub></b> SVM	<b>93.8</b>
	NN	97.0	98.1	87.8	<b>DNG<sub>P</sub></b> NN	90.2

features, while the TOP-based histograms represent only the frequency of features. Specifically, our DNG contains structural and motion information in each element of the vector, while TOP-based methods mainly contain spatial or temporal information in each bin.

## 3 EXPERIMENTS

We evaluated the proposed method for dynamic texture and facial expression recognition. We used support vector machines (SVMs) and a nearest neighbor classifier (NN) with Euclidean distance to evaluate the performance of the proposed methods. Given that SVM makes binary decisions, we achieved multi-class classification by adopting a one-against-one technique. We performed a grid-search on the hyper-parameters in a 10-fold cross-validation scheme in the training set for parameter selection. The parameter setting producing the best cross-validation accuracy was selected. Moreover, we used a second-order DNG descriptor for all experiments.

### 3.1 Dynamic texture

We evaluated the performance of the proposed method for dynamic texture recognition in two different databases: UCLA [5] and DynTex++ [6]. We used a grid of size  $1 \times 1 \times 1$ , performed a 50/50 split on the datasets for use as training and testing, and reported the average output over 20 trials. To select the best kernel,

we tested the histogram intersection (Inter.), Chi-square ( $\chi^2$ ), and RBF kernels in all the texture databases to evaluate our DNG descriptors (using Kirsch,  $DNG_K$ ; 3D Sobel,  $DNG_S$ ; and a nine-plane mask,  $DNG_P$ ; with second order descriptors), as shown in Table 1. In these scenarios, the best recognition was achieved by the RBF kernel, followed closely by the  $\chi^2$  kernel. Therefore, in the following experiments, we used the RBF kernel for the SVM. Additionally, we computed a histogram of the directions of the mask as a descriptor, instead of our DNG descriptor, to evaluate the contribution of the transitions. As shown in Table 1, the proposed method outperformed the histogram descriptor (Hist.), using SVM (with RBF kernel) and NN classifiers, on all the texture databases.

### 3.1.1 UCLA

The UCLA database [5] contains 50 dynamic texture classes, each with four grayscale video sequences captured from different viewpoints. All the samples we used were cropped from the videos, and each had a size of  $48 \times 48 \times 75$ . We evaluated UCLA using nine- and eight-class [6], [7] and shift-invariant-recognition (SIR) [1] breakdowns using a random partition of the classes, 50% for training and the remaining 50% for testing over 20 runs. The nine-class breakdown uses all the sequences grouped by type regardless of viewpoint. The eight-class excludes the “plants” class, since it contains too many sequences and thus may bias the results. Further, the SIR breakdown was created with shift-invariant textures by cropping the samples from the left and right portions of each video. In the latter, we performed comparisons between the left and right locations alone. This set up yields a two step test involving training with the left part and testing with the right one, and vice versa.

To evaluate our descriptors, we compared them against the dynamic fractal spectrum method (DFS) [27], DL-PEGASOS [6], a bag of dynamical systems (BoS) based method [7], [28], and a histogram of oriented gradients over nine planes of symmetry (HOG-NSP) method [11], which extracts the HOG feature over the cross-sections of the volume. Table 2(a) shows that the proposed methods perform better in all the breakdowns in comparison to previous methods. Moreover, we noted that the use of the 3D masks in the different spatiotemporal planes produced better result than the simple spatial response analysis, as  $DNG_P$  has an average improvement of 3% over  $DNG_K$  in all the breakdowns.

We show the confusion matrix for  $DNG_K$  and  $DNG_P$  in Table 3 in the SIR breakdown, as it is the more challenging breakdown. The greatest confusion occurs for  $DNG_K$  in the smoke sequences, due to the use of spatial information alone. For example, the spatial texture of smoke was confused with fire and plants, and boil and fountain were confused with plants. However, these issues are diminished by incorporating dynamic information through the use of the spatiotemporal masks

TABLE 3: Confusion matrix using SVM (RBF) in the SIR-UCLA.

(a) $DNG_K$									
(%)	boil	fire	flower	fount	plant	sea	smoke	water	wfalls
boil	43.75				50			6.25	
fire		75		18.75	6.25				
flower			66.67		33.33				
fount			7.5	45	42.5			2.5	2.5
plant	0.93		9.26	1.39	87.96			0.46	
sea						100			
smoke		25			50		25		
water								100	
wfalls									100

(b) $DNG_P$									
(%)	boil	fire	flower	fount	plant	sea	smoke	water	wfalls
boil	75				25				
fire		93.75							6.25
flower			29.17	8.33	62.5				
fount			2.5	87.5	10				
plant	0.93		6.48		92.59				
sea						100			
smoke		12.5					87.5		
water								100	
wfalls									100

in  $DNG_P$ . Nevertheless, any 3D mask does not guarantee the best results. For example, the diversity of textures proved to be too difficult for the 3D Sobel mask ( $DNG_S$ ), as shown in Table 2(a). Nevertheless, for the SIR breakdown of UCLA with both 3D masks,  $DNG_S$  and  $DNG_P$  have comparable recognition rates, but both showed a major confusion of flowers with plants.

Moreover, we also tested our methods with a nearest neighbor (NN) classifier to differentiate the contribution of the DNG descriptors and the SVM; the results are shown in Table 2(a). In that case,  $DNG_P$  is, on average, 2.6% better than HOG-NSP, and the other two DNG descriptors demonstrate general improvement compared to HOG-NSP. Nevertheless, the NN classifier has a lower recognition rate for the proposed methods than its SVM counterpart.

### 3.1.2 DynTex++

The DynTex++ dataset proposed by Ghanem and Ahuja [6] is a challenging dataset comprised of 36 classes of dynamic texture, each of which contains 100 sequences of a fixed size  $50 \times 50 \times 50$ . We used the same experimental settings as Ghanem and Ahuja [6] in the evaluation, that is, we use SVM as the classifier, train on 50% of the dataset, and test on the rest over 20 trials. We did not resize the sequences.

Similar to the experiment in the UCLA database, we compared against DFS [27], DL-PEGASOS [6], and HOG-NSP [11] and additionally included an LBP from three orthogonal planes (LBP-TOP) [8]. Table 2(b) shows the accuracy of all the methods, in which the proposed method  $DNG_P$  outperforms previous works by more than 3%. Moreover, we report the best accuracy for the HOG-NSP that relies on multiple kernel learning. However, when the HOG-NSP uses a simpler grid division framework (similar to that of our proposed method),

TABLE 4: FER accuracy (%) for (a) CK+ and (b) MMI.

(a)		(b)	
Method	(%)	Method	(%)
LBP-TOP [24]	90.8	B-LBP [13]	86.9
VLPQ [24]	91.4	CPL [31]	49.4
LPQ-TOP [24]	89.6	CSPL [31]	73.5
STLMBP [24]	92.6	LFEA [20]	94.1
CLM-SRI [21]	96.0	VTB [22]	95.0
<b>DNG<sub>K</sub></b>	<b>100</b>	<b>DNG<sub>K</sub></b>	<b>97.6</b>
<b>DNG<sub>S</sub></b>	<b>100</b>	<b>DNG<sub>S</sub></b>	<b>97.6</b>
<b>DNG<sub>P</sub></b>	<b>100</b>	<b>DNG<sub>P</sub></b>	<b>97.6</b>

its accuracy drastically decreases to 78.7%, so that even the use of NN in our methods demonstrates better performance. Furthermore, the DNG<sub>S</sub> method performs better than the 2D mask method, DNG<sub>K</sub>, as the Dytex++ database contains more structured textures in comparison to UCLA dataset. Therefore, our proposed DNG code scheme is more robust and accurate with a simpler extraction mechanism than previous methods.

### 3.2 Facial expressions

We performed experiments to evaluate the performance of the proposed algorithm under six- and seven-class facial expression recognition (FER). We tested our method in three different databases: Extended Cohn-Kanade (CK+) [29], MMI [30], and Oulu-CASIA [23] (using visual light and near-infrared datasets). Moreover, we cropped and normalized all the images to  $110 \times 100$  pixels, based on the ground truth positions of the eyes and mouth (when available) or using a face detector. We tested several partition combinations (using all permutations of 3, 5, 7, 9, and 11 parts for the spatial grid and 3, 5, and 7 parts for the temporal grid, as shown in Fig. 3) on the CK+ and MMI databases for our three masks, and reported the average values. Based on this experiment, we chose to use the best combination  $11 \times 3 \times 7$  for the rest of the experiments, and an illustration of the grid is shown in Fig. 3(d). For all the databases, we performed every experiment 10 times and reported the average values after randomly choosing 90% of the database for training and the rest for testing using SVM as a classifier.

#### 3.2.1 Extended Cohn-Kanade

We used the extended Cohn-Kanade database (CK+) [29], which includes 325 sequences from 118 subjects demonstrating seven basic expressions (happiness, sadness, surprise, anger, disgust, fear, and contempt). To evaluate our descriptor, we compared it against several methods, such as LBP-TOP [8], volume local phase quantization (VLPQ) [9], local phase quantization from three orthogonal planes (LPQ-TOP) [9], a spatiotemporal local monogenic binary pattern method fusing real and imaginary data (STLMBP) [24], and an extended constrained-local-model algorithm with action unit normalization

TABLE 5: Confusion matrix of DNG<sub>P</sub> on MMI.

(%)	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	98.21					1.79
Disgust		95.65	4.35			
Fear			97.92			2.08
Happiness				100		
Sadness		5.36			94.64	
Surprise			1.43			98.57

(CLM-SRI) [21]. Table 4(a) shows that our DNG-based methods outperformed all the other methods. Moreover, the lower accuracy of the methods that use three orthogonal planes for the dynamic information representation may be due to their combination of the spatiotemporal information in the histogram level alone. Instead, the mixture of spatial and temporal data in early stages of the coding process enriches the sequence descriptor, as the spatiotemporal patterns remain coupled rather than being split due to the cross-sectioning process in TOP-based methods. Furthermore, Guo *et al.* [20] reported an accuracy of 97.2% for their longitudinal atlases on the non-extended Cohn-Kanade database, and Ji and Idrissi [22] reported a 97.3% accuracy for their LBP-based method. Our methods outperformed these on the more complex version of the database, recognizing all seven expressions instead of only six.

#### 3.2.2 MMI

Moreover, we tested the expression recognition problem on the MMI face database [30]. In our experiments, we used Part II of the database, which comprises 238 sequences of 28 subjects (sessions 1767 to 2004) where all expressions (anger, disgust, fear, happiness, sadness, and surprise) were recorded twice.

We compared our proposed methods against other recent studies, such as a boosted LBP method (B-LBP) [13], several patch-based approaches, common patches (CPL), and common and specific patches (CSPL) [31], longitudinal facial expression atlases (LFEA) [20], and an LBP-based vertical time backward method (VTB) [22]. Table 4(b) shows that DNG outperforms previous methods. Additionally, our method is not boosted in any way, unlike these other methods. Moreover, we use a wide variety of images and all expressions to evaluate our performance, while Shan *et al.* [13] used a reduced set. The worst confusion was observed for the DNG<sub>P</sub> descriptor for disgust and sadness expressions, as shown in Table 5.

#### 3.2.3 Oulu-CASIA

The Oulu-CASIA facial expression database [23] contains six expressions (surprise, happiness, sadness, anger, fear and disgust) from 80 people captured by near-infrared (NIR) and visible light (VIS) cameras; it is comprised of 1440 and 1438 image sequences from VIS and NIR datasets, respectively. All expressions were captured in

$c \backslash r$	3	5	7	9	11
3	96.83	96.98	97.1	97.16	97.26
5	97.4	98.05	97.63	97.73	97.65
7	98.37	98.35	98.03	97.66	97.95
9	98.38	98.4	97.88	98.03	97.93
11	98.43	98.63	98.13	98.08	97.95

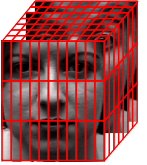
(a)  $t = 3$

$c \backslash r$	3	5	7	9	11
3	92.81	96.65	96.95	97.32	97.54
5	97.18	97.65	97.58	97.58	97.36
7	98.03	98.00	97.83	97.86	97.83
9	98.02	98.03	97.76	97.76	97.71
11	98.61	98.31	97.96	98.11	97.86

(b)  $t = 5$

$c \backslash r$	3	5	7	9	11
3	93.47	96.72	97.24	97.45	97.66
5	97.14	97.97	97.71	98.03	97.93
7	98.14	98.16	97.96	98.11	97.98
9	98.78	98.46	98.06	98.11	97.98
11	98.81	98.41	97.83	98.01	97.86

(c)  $t = 7$



(d)

Fig. 3: (a) (b) (c) Average FER accuracy (%) using different grid resolutions ( $r$ ,  $c$ , and  $t$  represent row, columns, and time, respectively). (d) Illustration of the best grid in a facial image sequence.

TABLE 6: FER accuracy (%) for the Oulu-CASIA (VIS) database.

Method	Normal (%)	Weak (%)	Dark (%)
LBP-TOP [24]	76.2	65.3	56.3
LFEA [20]	75.5	61.8	57.7
STLMBP [24]	79.9	64.6	62.0
<b>DNG<sub>K</sub></b>	96.4	97.7	93.3
<b>DNG<sub>S</sub></b>	<b>98.5</b>	<b>99.0</b>	<b>98.6</b>
<b>DNG<sub>P</sub></b>	97.8	98.9	<b>98.6</b>

TABLE 7: Confusion matrix of DNG<sub>S</sub> on Oulu-CASIA (VIS) dark.

(%)	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	98.75	1.25				
Disgust	2.53	97.47				
Fear			97.81	0.63	1.25	0.31
Happiness		0.63		98.73	0.63	
Sadness				0.63	99.38	
Surprise						99.38

three different illumination conditions (normal, weak, and dark).

We compared our algorithm against the best version of LBP-TOP proposed by Zhao *et al.* [23], in which they use a sparse representation classifier, LFEA [20], and STLMBP [24]. Table 6 shows the results in the three different illumination conditions, in which the proposed method considerably outperforms the others. Interestingly, DNG<sub>S</sub> and DNG<sub>P</sub> present higher accuracy for weak illumination than for the normal illumination. This may due to non-homogeneous illumination throughout the datasets. In the dark setting (Table 7) DNG<sub>S</sub> demonstrated the greatest confusion among all our descriptors.

We also experimented on the NIR part of the Oulu-CASIA database and compared our DNG-based methods against LBP-TOP [23] and STLMBP [24]. Table 8 shows that DNG<sub>P</sub> surpassed the other methods. Moreover, our proposed coding scheme is more stable in comparison to other methods with similar recognition accuracy in the three datasets, around 97%, 98%, and 98% for the 2D and 3D masks, respectively. This stability results from the constancy of the near-infrared images in the presence of light changes. Furthermore, Table 9 shows the confusion matrices of the NIR images, which demonstrated better recognition accuracy per expression compared with the VIS counterpart. In the three datasets, the fear and sadness expressions were most often confused.

## 4 CONCLUSION

In this paper, we introduced a new descriptor for image sequences that jointly models the motion and spatial

TABLE 8: FER accuracy (%) for Oulu-CASIA (NIR) database.

Method	Normal (%)	Weak (%)	Dark (%)
LBP-TOP [24]	78.6	73.3	70.4
STLMBP [24]	78.7	70.4	72.3
DNG <sub>K</sub>	97.0	97.5	97.3
<b>DNG<sub>S</sub></b>	<b>98.5</b>	<b>99.2</b>	98.6
<b>DNG<sub>P</sub></b>	<b>98.5</b>	<b>99.2</b>	<b>98.7</b>

TABLE 9: Confusion matrix of DNG<sub>P</sub> on Oulu-CASIA (NIR).

(a) normal

(%)	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	97.5	1.25	1.25			
Disgust		99.38			0.63	
Fear			98.1	0.63	1.27	
Happiness			0.63	99.38		
Sadness	0.63	0.63		0.63	97.47	0.63
Surprise			0.63			99.38

(b) weak

(%)	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	98.75	0.63	0.63			
Disgust	1.25	98.13			0.63	
Fear			99.37		0.63	
Happiness			0.63	99.38		
Sadness			0.63		99.37	
Surprise						100

(c) dark

(%)	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Anger	100					
Disgust	1.9	96.2			1.9	
Fear			99.38		0.63	
Happiness				98.73	0.63	0.63
Sadness	0.63		0.63		98.75	
Surprise					0.63	99.38

structure of dynamic patterns. The main advantage of the proposed Directional Number Transitional Graph descriptor is the extraction of low-level features that are later combined using a two-layer descriptor including a graph that models the intrinsic motion of the lower features and a spatiotemporal grid that maintains the spatiotemporal relations between the created regions. Moreover, we explored two different approaches to represent the lower features through directional numbers: (1) using pure spatial information (2D compass mask) and (2) extracting spatiotemporal directional information as the planar response in nine principal directions (3D compass mask). Our results show that the inclusion of motion and spatial information in early stages of the coding process enhances the recognition rates of dynamic patterns compared to using cross-sections and subsequent mixing mechanisms.

## REFERENCES

- [1] K. Derpanis and R. Wildes, "Dynamic texture recognition based on distributions of spacetime oriented structure," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010, pp. 191–198.
- [2] D. Chetverikov and R. Péteri, "A brief survey of dynamic texture description and recognition," in *Computer Recognition Systems*, ser. Advances in Soft Computing, M. Kurzyński, E. Puchala, and A. Woźniak, Michałand żolnierek, Eds. Springer Berlin Heidelberg, 2005, vol. 30, pp. 17–26.
- [3] Z. Lu, W. Xie, J. Pei, and J. Huang, "Dynamic texture recognition by spatio-temporal multiresolution histograms," in *Application of Computer Vision, 2005. WACV/MOTIONS '05 Volume 1. Seventh IEEE Workshops on*, vol. 2, Jan. 2005, pp. 241–246.
- [4] R. Péteri and D. Chetverikov, "Dynamic texture recognition using normal flow and texture regularity," in *Proc. Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA 2005)*. Springer, 2005, pp. 223–230.
- [5] P. Saisan, G. Doretto, Y. N. Wu, and S. Soatto, "Dynamic texture recognition," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 2, 2001, pp. II–58–II–63.
- [6] B. Ghanem and N. Ahuja, "Maximum margin distance learning for dynamic texture recognition," in *Computer Vision – ECCV 2010*, ser. Lecture Notes in Computer Science, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Springer Berlin Heidelberg, 2010, vol. 6312, pp. 223–236.
- [7] A. Ravichandran, R. Chaudhry, and R. Vidal, "View-invariant dynamic texture recognition using a bag of dynamical systems," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, Jun. 2009, pp. 1651–1657.
- [8] G. Zhao and M. Pietikäinen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.
- [9] J. Päiväranta, E. Rahtu, and J. Heikkilä, "Volume local phase quantization for blur-insensitive dynamic texture classification," in *Image Analysis*, ser. Lecture Notes in Computer Science, A. Heyden and F. Kahl, Eds. Springer Berlin Heidelberg, 2011, vol. 6688, pp. 360–369.
- [10] G. Zhao, T. Ahonen, J. Matas, and M. Pietikäinen, "Rotation-invariant image and video description with local binary pattern features," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1465–1477, 2012.
- [11] E. Norouzzehad, M. Harandi, A. Bigdeli, M. Baktash, A. Postula, and B. Lovell, "Directional space-time oriented gradients for 3d visual pattern analysis," in *Computer Vision – ECCV 2012*, ser. Lecture Notes in Computer Science, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Springer Berlin Heidelberg, 2012, vol. 7574, pp. 736–749.
- [12] W. Zhang, S. Shan, L. Qing, X. Chen, and W. Gao, "Are gabor phases really useless for face recognition?" *Pattern Analysis and Applications*, vol. 12, pp. 301–307, 2009.
- [13] C. F. Shan, S. G. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.
- [14] M. Kabir, T. Jabid, and O. Chae, "A local directional pattern variance (LDPv) based face descriptor for human facial expression recognition," in *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*, Sep. 2010, pp. 526–532.
- [15] A. Ramirez Rivera, J. Rojas Castillo, and O. Chae, "Recognition of face expressions using local principal texture pattern," in *Image Processing, 2012. ICIP 2012. IEEE International Conference on*, Sep. 2012.
- [16] —, "Local gaussian directional pattern for face recognition," in *International Conference on Pattern Recognition (ICPR)*, Nov. 2012, pp. 1000–1003.
- [17] —, "Local directional number pattern for face analysis: Face and expression recognition," *IEEE Trans. Image Process.*, 2012.
- [18] J. Rojas Castillo, A. Ramirez Rivera, and O. Chae, "Facial expression recognition based on local sign directional pattern," in *Image Processing, 2012. ICIP 2012. IEEE International Conference on*, Sep. 2012.
- [19] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.
- [20] Y. Guo, G. Zhao, and M. Pietikäinen, "Dynamic facial expression recognition using longitudinal facial expression atlases," in *Computer Vision – ECCV 2012*, ser. Lecture Notes in Computer Science, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds. Springer Berlin Heidelberg, 2012, pp. 631–644.
- [21] L. Jeni, D. Takacs, and A. Lorincz, "High quality facial expression recognition in video streams using shape related information only," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, Nov. 2011, pp. 2168–2174.
- [22] Y. Ji and K. Idrissi, "Automatic facial expression recognition based on spatiotemporal descriptors," *Pattern Recognition Letters*, vol. 33, no. 10, pp. 1373–1380, 2012.
- [23] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011.
- [24] X. Huang, G. Zhao, W. Zheng, and M. Pietikäinen, "Spatiotemporal local monogenic binary patterns for facial expression recognition," *IEEE Signal Process. Lett.*, vol. 19, no. 5, pp. 243–246, May 2012.
- [25] R. A. Kirsch, "Computer determination of the constituent structure of biological images," *Computers & Biomedical Research*, pp. 315–328, 1970.
- [26] L. Jetto, G. Orlando, and A. Sanfilippo, "The edge point detection problem in image sequences: Definition and comparative evaluation of some 3d edge detecting schemes," in *7th IEEE Mediterranean Conference on Control and Automation (MED '99)*, Jun. 1999, pp. 2161–2171.
- [27] Y. Xu, Y. Quan, H. Ling, and H. Ji, "Dynamic texture classification using dynamic fractal analysis," in *Computer Vision (ICCV), 2011 IEEE International Conference on*, Nov. 2011, pp. 1219–1226.
- [28] A. Ravichandran, R. Chaudhry, and R. Vidal, "Categorizing dynamic textures using a bag of dynamical systems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 2, pp. 342–353, 2013.
- [29] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "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*, Jun. 2010, pp. 94–101.
- [30] M. F. Valstar and M. Pantic, "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, May 2010, pp. 65–70.
- [31] L. Zhong, Q. Liuz, P. Yangy, B. Liuy, J. Huangx, and D. N. Metaxasy, "Learning active facial patches for expression analysis," in *Computer Vision and Pattern Recognition, 2012. Proceedings. IEEE Conference on*, 2012.



**Adin Ramirez Rivera** (S'12, M'14) received his B.Sc. degrees in Computer Engineering from San Carlos University (USAC), Guatemala in 2009. He completed his M.Sc. and Ph.D. degree in Computer Engineering from Kyung Hee University, South Korea in 2013. He is currently an Assistant Professor at Escuela de Informática y Telecomunicaciones, Facultad de Ingeniería, Universidad Diego Portales, Chile. His research interests are image enhancement, object detection, expression recognition, and pattern recognition.



**Oksam Chae** (M'92) received his B.Sc. degree in Electronics Engineering from Inha University, South Korea in 1977. He completed his M.S. and Ph.D. degrees in Electrical and Computer Engineering from Oklahoma State University, USA in 1982 and 1986, respectively. From 1986 to 1988, he worked as a Research Engineer at Texas Instruments Image Processing Laboratory, USA. Since 1988, he has been a Professor in the Department of Computer Engineering, Kyung Hee University, South Korea. His research interests include multimedia data processing environment, intrusion detection, sensor networks, and dental image processing.