Cleveland State University
EngagedScholarship@CSU

Electrical Engineering & Computer Science Faculty Publications

Electrical Engineering & Computer Science Department

2019

High Order Volumetric Directional Pattern for Video-Based Face Recognition

Almabrok Essa
Cleveland State University, a.essa@csuohio.edu

Vijayan Asari
University of Dayton

Follow this and additional works at: https://engagedscholarship.csuohio.edu/enece_facpub

 Part of the Electrical and Computer Engineering Commons

How does access to this work benefit you? Let us know!

Repository Citation

Essa, Almabrok and Asari, Vijayan, "High Order Volumetric Directional Pattern for Video-Based Face Recognition" (2019). *Electrical Engineering & Computer Science Faculty Publications*. 457.
https://engagedscholarship.csuohio.edu/enece_facpub/457

This Article is brought to you for free and open access by the Electrical Engineering & Computer Science Department at EngagedScholarship@CSU. It has been accepted for inclusion in Electrical Engineering & Computer Science Faculty Publications by an authorized administrator of EngagedScholarship@CSU. For more information, please contact library.es@csuohio.edu.

Hindawi
Mathematical Problems in Engineering
Volume 2019, Article ID 6798750, 10 pages
https://doi.org/10.1155/2019/6798750

Research Article

High Order Volumetric Directional Pattern for Video-Based Face Recognition

Almabrok Essa[1] **and Vijayan Asari**[2]

[1]*Department of Electrical Engineering and Computer Science, Cleveland State University, 2121 Euclid Ave, 44115 Cleveland, OH, USA*
[2]*University of Dayton, Dayton, OH, USA*

Correspondence should be addressed to Almabrok Essa; a.essa@csuohio.edu

Received 28 November 2018; Revised 5 April 2019; Accepted 2 June 2019; Published 25 June 2019

Academic Editor: George A. Papakostas

Copyright © 2019 Almabrok Essa and Vijayan Asari. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Describing the dynamic textures has attracted growing attention in the field of computer vision and pattern recognition. In this paper, a novel approach for recognizing dynamic textures, namely, high order volumetric directional pattern (HOVDP), is proposed. It is an extension of the volumetric directional pattern (VDP) which extracts and fuses the temporal information (dynamic features) from three consecutive frames. HOVDP combines the movement and appearance features together considering the n^{th} order volumetric directional variation patterns of all neighboring pixels from three consecutive frames. In experiments with two challenging video face databases, YouTube Celebrities and Honda/UCSD, HOVDP clearly outperformed a set of state-of-the-art approaches.

1. Introduction

The texture of objects in digital images can be generally categorized into two main types, static texture and dynamic texture, which is an extension of texture in the temporal domain. Local feature detection and description have gained much attention in recent years since photometric descriptors computed for regions of interest have proven to be very successful in many computer vision applications. In the context of texture (feature) analysis methods, there are two common types of techniques: (1) the structural approaches, where the image texture is considered as a repetition of some primitives with a specific rule of placement, and (2) the statistical methods. The stochastic properties of the spatial distribution of gray levels in an image are characterized by the gray tone cooccurrence matrix. A set of textural features derived from the cooccurrence matrix is widely used to extract textural information from digital images [1].

Face recognition (FR) is one of the most suitable technologies that has been spread in several applications such as biometric systems, access control and information security systems, surveillance systems, content-based video retrieval systems, credit-card verification systems, and more generally image understanding. FR is a biometric approach that employs automated methods to verify or recognize the identity of a living person based on their physiological characteristics. The key of each face recognition system is the feature extractors, which should be distinct and stable under different conditions. FR system can generally be categorized into one of the two main scenarios based on the characteristics of the images to be matched, such as still-image-based (still-to-still) FR [2–4] or video-based (video-to-video) FR. Also it could be a video-to-still-image-based face recognition system [5].

Dynamic texture (DT) or temporal texture is a texture with motion that includes the class of video sequences, which offers some stationary properties in time. Recently researchers start to investigate the domain of video, where the problem of face recognition becomes more challenging due to pose variations, different facial expressions, illumination changes, occlusions, and so on. However, DT provides many samples of the same person, thus providing the opportunity to convert many weak examples into a strong prediction of the identity. Zhao and Pietikäinen introduced an extended version of LBP named volume local binary patterns (VLBP) for video-based facial expression recognition [6]. They claim

that the features extracted in a small local neighborhood of the volume can be boosted by combining the motion and appearance. These features are insensitive to translation and rotation. However, there are some illumination limitations since this method deals with the small local neighborhood of a pixel as well as utilizing the image intensity directly. While there are other methods for still and video-based face recognition procedure that do not rely on the image features, they depend on the sparse representation based categorization strategy, which considers the local sparsity identification from sparse coding coefficients [7, 8]. Zheng et al. [9] recently introduced a full system for unconstrained video-based face recognition, which is composed of face/fiducial detection, face association, and face recognition.

Nowadays, manifold features (linear subspaces), if the features lie in Euclidean spaces, have proven a powerful representation for video-based face recognition. Huang et al. [10] recently introduced a new method called projection metric learning on Grassmann manifold (PML), which is combined with Grassmannian graph-embedding discriminant analysis (GGDA) [11]. In this technique, each video sequence can be treated as a set of face images without considering the temporal information. It serves as both a metric learning and a dimensionality reduction method for the Grassmann manifold to map the manifold to a reproducing kernel Hilbert space (RKHS). Although kernel-based methods have been successfully used in many computer vision applications, poor choice of kernels can often result in degraded classification performance [12], especially when the data lies in non-Euclidean spaces. Paivarinta et al. [13] introduced a blur-insensitive descriptor for dynamic textures, named volume local phase quantization (VLPQ). It is based on binary encoding of the phase information of the local Fourier transform. In this technique, each video sequence is processed to provide one feature vector.

In this paper, we introduce a new video-based facial feature extractor, named high order volumetric directional pattern (HOVDP). HOVDP is a histogram bin-based code assigned to each pixel of an input frame, which can be calculated by fusing twenty-four edge responses from three consecutive frames. These gradient values are detected using Kirsch masks in eight different directions. If there is any change (dynamic changes) in any relative gradient response from the corresponding frames at any direction, it would be detected and added to the features vector. Unlike the conventional VDP operator that encodes the volumetric directional information in the small 3×3 local neighborhood of a pixel of each three consecutive frames, HOVDP extracts n^{th} order volumetric information by encoding various distinctive spatial relationships from each neighborhood layer of a pixel in the pyramidal multistructure way and then concatenating the feature vector of each neighborhood layer to form the final HOVDP-feature vector.

The remainder of the paper is organized as follows. Section 2 introduces the observation model and related work. The proposed HOVDP algorithm is presented in Section 3. Experimental results and analysis are presented in Section 4. Finally, we offer conclusions in Section 5.

2. Related Work

In this section, we present the observation of capturing any small changes of the face textures and merging the movement and appearance features together. Therefore, we deeply explain the essentials of the proposed technique which is our previous work, named volumetric directional pattern (VDP) [14, 15]. The main goal of the VDP is extracting and fusing the temporal information (dynamic features) from three consecutive frames which are distinct under multiple poses and facial expressions variations. Given a video as input and a gallery of videos, we perform face recognition process throughout the whole video clip. Firstly, we detect faces using Viola-Jones's face detector [16]. Then for each frame we extract and combine the dynamic features of its two neighborhood frames. Then a histogram is built for each frame. These histograms are concatenated to form the final VDP-feature vector, similar to the gallery videos.

2.1. Volumetric Directional Pattern. Volumetric directional pattern (VDP) is a gray-scale pattern that characterizes and fuses the temporal structure (dynamic information) of three consecutive frames [14, 15]. VDP has been developed to merge the movement and appearance features together. It is a twenty-four-bit binary code assigned to each pixel of an input frame, which can be calculated by comparing the relative edge response value of a particular pixel from three consecutive frames in different directions by using Kirsch masks in eight different orientations centered on its own position for one frame and the corresponding positions of the other two frames. Kirsch mask is a first derivate filter which is used to detect edges in all eight directions of a compass considering all eight neighbors [17]. Specifically, it takes a single mask, denoted as $M_i(x, y)$ for $i = 0, 1, \ldots, 7$, and rotates it in 45° increments through all 8 compass directions as follows:

$$\begin{matrix}
\begin{bmatrix} -3 & -3 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & 5 \end{bmatrix} & \begin{bmatrix} -3 & 5 & 5 \\ -3 & 0 & 5 \\ -3 & -3 & -3 \end{bmatrix} \\
\text{East } (M_0) & \text{North East } (M_1) \\
\begin{bmatrix} 5 & 5 & 5 \\ -3 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix} & \begin{bmatrix} 5 & 5 & -3 \\ 5 & 0 & -3 \\ -3 & -3 & -3 \end{bmatrix} \\
\text{North } (M_2) & \text{North West } (M_3) \\
\begin{bmatrix} 5 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & -3 & -3 \end{bmatrix} & \begin{bmatrix} -3 & -3 & -3 \\ 5 & 0 & -3 \\ 5 & 5 & -3 \end{bmatrix} \\
\text{West } (M_4) & \text{South West } (M_5) \\
\begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & -3 \\ 5 & 5 & 5 \end{bmatrix} & \begin{bmatrix} -3 & -3 & -3 \\ -3 & 0 & 5 \\ -3 & 5 & 5 \end{bmatrix} \\
\text{South } (M_6) & \text{South East } (M_7)
\end{matrix} \tag{1}$$

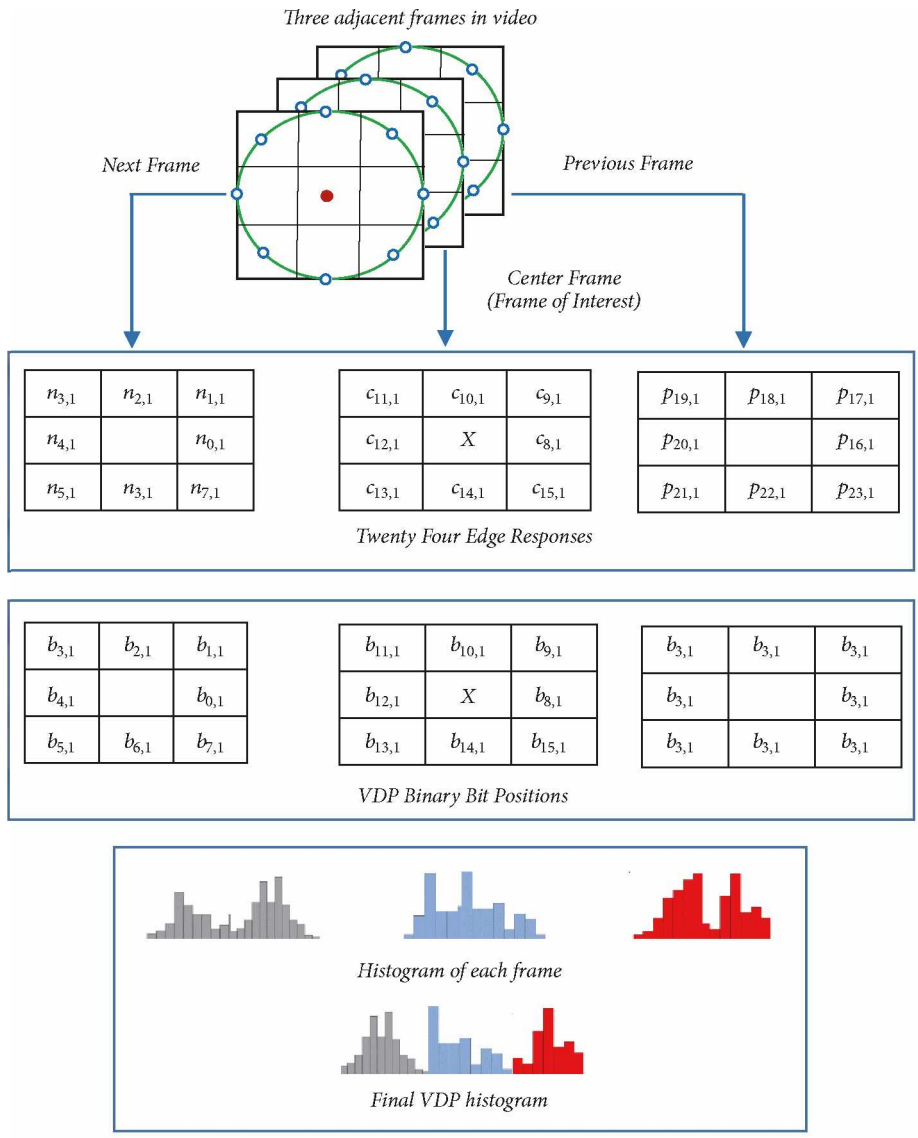


Figure 1: The twenty-four edge responses with their VDP binary bit positions and the final VDP-feature vector.

Given a central pixel in the middle (center) frame of three consecutive frames, the eight different directional edge response values c_i for $i = 8, 9, \ldots, 15$ are used to create an eight bit binary number which can describe the edge response pattern of each pixel in the center frame (frame of interest). Meanwhile, the eight different edge response values p_i for $i = 16, 17, \ldots, 23$ and n_i for $i = 0, 1, \ldots, 7$ are used to create an eight-bit binary number each, which can describe the edge response pattern of each pixel in the previous frame and next frame, respectively. Figure 1 shows the twenty-four edge responses and their corresponding bit binary positions, as well as the fusing strategy of this 24-bit code. The twenty-four different directional edge response values for each pixel location can be computed by

$$p = \sum_{0}^{7} dot\left(I_p, M_i\right), \tag{2}$$

$$c = \sum_{0}^{7} dot\left(I_c, M_i\right), \tag{3}$$

and

$$n = \sum_{0}^{7} dot\left(I_n, M_i\right) \tag{4}$$

where $dot(\cdot)$ represents the dot product operation, M_i is the mask, and I_p, I_c, and I_n are the 3×3 neighbors of each pixel of previous, center, and next frames, respectively. p, c, and n are the spatiotemporal directional response values of the first layer for the previous, center, and next frames, respectively.

In order to generate the VDP-feature vector, we need to know the t most prominent directional bits for all three consecutive frames. These t bits are set to 1 and the rest of 8-bit VDP pattern of each frame are set to 0. Then a binary code is formed to each pixel from each frame, which will

be mapped to its own bin to build a histogram. Finally, we concatenate these three histograms of these three consecutive frames to obtain the final VDP-feature vector, which is the descriptor for each center frame (frame of interest) that we used to recognize the face image by the help of a classifier. The final VDP code can be derived by

$$VDP = \sum_{i=0}^{7} f(n_i - n_t) \times 2^i \parallel \sum_{i=8}^{15} f(c_i - c_t) \times 2^{i-8} \parallel \sum_{i=16}^{23} f(p_i - p_t) \times 2^{i-16} \quad (5)$$

and

$$f(x) = \begin{cases} 1 & \text{if } x \geq 0 \\ 0 & \text{if } x < 0 \end{cases} \quad (6)$$

where the vertical bars ($\parallel$) represent the concatenation process of the histograms.

3. Method

Derived from a general definition of texture in a local neighborhood, the conventional VDP or the first order volumetric directional pattern encodes the directional information in the small 3 × 3 local neighborhood of a pixel of each three consecutive frames, which may fail to extract detailed information especially during changes in data collection environments. Therefore, we proposed an improved version of VDP to tackle this problem by calculating the n^{th} order volumetric directional variation patterns, namely, high order volumetric directional pattern (HOVDP). The proposed HOVDP can capture more detailed discriminative information than the traditional VDP. Unlike the VDP operator, the proposed HOVDP technique extracts n^{th} order volumetric information by encoding various distinctive spatial relationships from each neighborhood layer of a pixel in the pyramidal multistructure way and then concatenating the feature vector of each neighborhood layer to form the final HOVDP-feature vector. Several observations can be made for HOVDP:

(i) Under the proposed framework, the original VDP is a special case of HOVDP, which simply calculates the 1^{st} order volumetric directional information in the local neighborhood of a pixel.

(ii) The relation between the neighbor layers and the pixel under consideration could be easily weighted in HOVDP based on the distance between each layer and the central pixel. Because of that, the pixels within the closest layer to the central pixel has more weight than the others.

(iii) Due to the same format and feature length of different order HOVDP, they can be readily fused, and the accuracy of the face recognition can be significantly improved after the fusion.

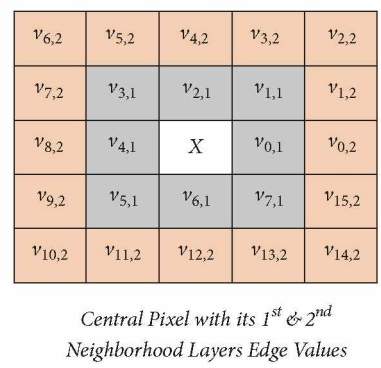


Figure 2: Magnitude values of two neighborhood layers.

3.1. High Order Volumetric Directional Pattern. The proposed high order volumetric directional pattern (HOVDP) technique is an oriented and multiscale volumetric directional descriptor that is able to extract and fuse the information of multiple frames, temporal (dynamic) information, and multiple poses and expressions of faces in input video to produce strong feature vectors. Given a central pixel in the middle (center) frame of three consecutive frames, to calculate the second order volumetric directional pattern (VDP_2) we first compute the first order (VDP_1), which is exactly the same as the original VDP by using the (2)–(6). Then,

$$n_1 = n, \quad (7)$$

$$c_1 = c, \quad (8)$$

$$p_1 = p, \quad (9)$$

and

$$VDP_1 = VDP \quad (10)$$

where p_1, c_1, and n_1 are an eight-bit binary number that describes the edge response pattern of each pixel of the first layer in the previous frame, center frame, and next frames, respectively.

To make our calculation simple and easy to compute the high order relevant edge values, let us assume v_i the magnitude values for each layer separately after convolving the input image $I(x, y)$ with Kirsch masks in eight different directions $M_i(x, y)$ for $i = 0, 1, \ldots, 7$, which can be seen in Figure 2. Then the directional edge values for the particular layer can be found as

$$c_1 = v_{i,1}, \quad (11)$$

$$c_2 = \frac{1}{3} \sum_{j=-1}^{j=1} v_{g,2}, \quad (12)$$

and

$$g = \text{mod}(2i + j, 16), \quad \text{for } i = 0, 1, \ldots, 7 \quad (13)$$

where c_1 and c_2 are the eight different directional edge response values of the first and second neighborhood layers,

respectively. $v_{g,2}$ is the magnitude value after convolving the input image with Kirsch kernels, the subscripts g and 2 are the number of surrounding pixels of each direction i and the second neighborhood layer (second order), respectively, and (mod) is the modulo operation which is used to maintain the circularly neighborhood configuration.

Based on the observation that every corner or edge has high response values in particular directions, we are interested to know t the most prominent directional bits for all three consecutive frames in order to generate the feature vector of each neighborhood layer. These t bits are set to 1 and the rest of 8-bit pattern in each layer from each frame are set to 0. Then a binary code is formed to each pixel in each layer from each frame, which will be mapped to its own bin to build a histogram for that particular layer of each frame. The volumetric directional pattern of each pixel position in the second neighbor layer can be formed as

$$VDP_2 = \sum_{i=0}^{7} f\left(n_{i,2} - n_{t,2}\right) \times 2^i \parallel \sum_{i=8}^{15} f\left(c_{i,2} - c_{t,2}\right) \times 2^{i-8} \parallel \sum_{i=16}^{23} f\left(p_{i,2} - p_{t,2}\right) \times 2^{i-16} \tag{14}$$

where the thresholding function $f(x)$ can be defined as in (6).

After identifying the volumetric directional pattern of each pixel in each neighborhood layer from each frame (VDP_1 for the first layer and VDP_2 for the second layer), a histogram of 256 bins is built to represent all the distinguishing features of each neighbor layer from each frame separately, which means each layer will provide a feature vector of 768-bin size. Then to obtain the final 2^{nd} order VDP-feature vector, which is the descriptor for each center frame (frame of interest), we concatenate these histograms starting from the same layer order one by one, which can be seen in Figure 3. The 2^{nd} order VDP-feature vector size would be 1536 bins (768×2). Therefore, the feature vector size for n^{th} order is $768 \times n$.

In a general formulation, the n^{th} order volumetric directional pattern (VDP_n) of each pixel position in each neighbor layer from each frame can be defined as

$$VDP_n = \sum_{i=0}^{7} f\left(n_{i,n} - n_{t,n}\right) \times 2^i \parallel \sum_{i=8}^{15} f\left(c_{i,n} - c_{t,n}\right) \times 2^{i-8} \parallel \sum_{i=16}^{23} f\left(p_{i,n} - p_{t,n}\right) \times 2^{i-16} \tag{15}$$

where the subscript $n = 1, 2, \ldots$ is the volumetric directional pattern order (the number of neighborhood layers that has been used for the calculation process); $n_{t,n}$, $c_{t,n}$, and $p_{t,n}$ are the t^{th} most significant directional responses of each neighboring layer n from next frame, center frame, and previous frame, respectively; and $f(x)$ is the thresholding function that can be defined as in (6). $n_{i,n}$, $c_{i,n}$, and $p_{i,n}$ are the eight different directional edge response values of each neighborhood layer n from next frame, center frame, and previous frame, respectively, which can be computed as

$$c_n = \frac{1}{2n-1} \sum_{j=-n+1}^{n-1} v_{g,n} \tag{16}$$

and

$$g = \mathrm{mod}\left(ni + j, 8n\right) \quad for\ i = 0, 1, \ldots, 7. \tag{17}$$

The same procedure is applied to find the eight different directional edge response values of next frame n_n and previous p_n.

4. Results and Discussion

To evaluate the robustness of the introduced method in illumination, pose, and expression variations, we tested it on two publicly available datasets, namely, YouTube Celebrities dataset [18] and Honda/UCSD database [19, 20]. All the face images in this work were detected by using the Viola and Jones's face detector [16]. After manually removing the false detection, all the detected face images were resized to 64×64, and then the spatiotemporal information was extracted using the proposed high order VDP technique. When it comes to the face recognition process, we represent the face using a HOVDP-feature histogram. The objective is to compare the encoded feature vector from one frame with all other candidate feature vectors using two well-known classifiers. The first one is support vector machine (SVM) classifier (we used LibSVM) [21], and the second one is a k-nearest neighbors (KNN) classifier. The corresponding face of the HOVDP-feature vector with the lowest measured value indicates the match found.

Two different experiments are conducted for each database to verify the effectiveness and efficiency of the proposed HOVDP framework. The first one explores the effectiveness of different volumetric directional order (different neighborhood layers for each pixel of the image) as changing the number of the most prominent response values $\{t = 2, 3, \ldots, 6\}$. The second one evaluates the effectiveness of the proposed HOVDP by comparing it with four popular video-based face recognition techniques as well as with our conventional VDP. To avoid any bias, we randomly selected the data for training and testing.

Considering computational efficiency, we observe that HOVDP requires longer execution time compared to the VLPQ method. For example, during the processing of one video clip that consists 248 frames, HOVDP takes around 2 minutes, while VLPQ requires around 2 seconds. The reason is that HOVDP computes features using every three adjacent frames for all pixels, which increases the feature dimension by adding each neighborhood layer; in contrast, the abovementioned competing VLPQ method computes the features based on Fourier transform estimation, which is performed locally using short-term Fourier transform using 1D convolutions, and the convolutions are computed using only valid areas, i.e., areas that can be computed without zero-padding. The convolutions that occur multiple times in the process are calculated only once [13]. The computing platform is an Intel Core i5 2.27-GHz machine with 4 GB of RAM, and all implementations are performed on MATLAB-2016a.

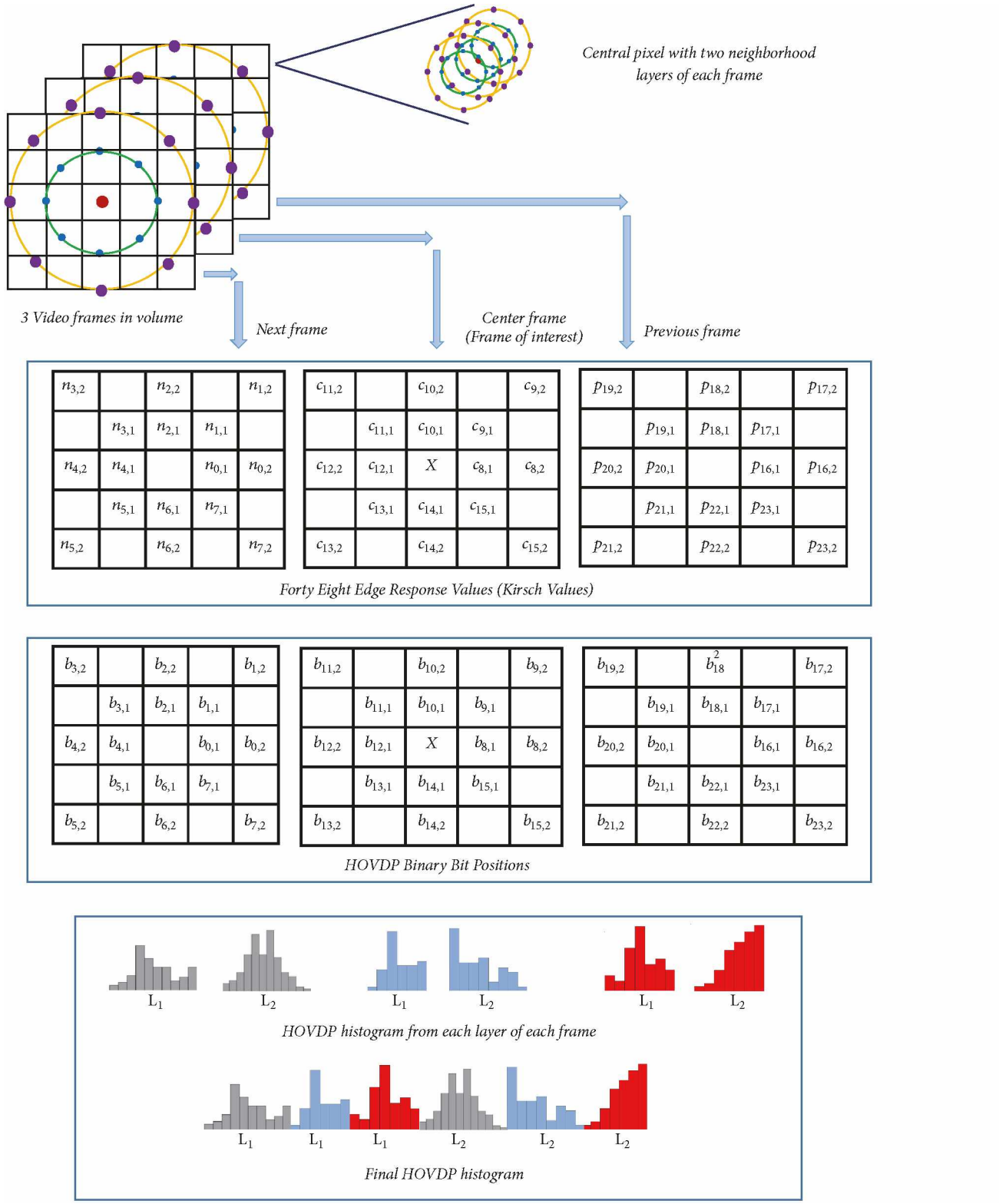


Figure 3: Procedure of 2^{nd} order VDP. L_1 and L_2 are the histograms of first and second layers, respectively, for each frame.

Table 1: Recognition rates using all video frames as changing the threshold t and the proposed approach order on YouTube celebrities database.

Descriptor	Recognition Accuracy (%) Using SVM Classifier				
Order	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$
1^{st} Order (VDP)	86.50 %	86.46 %	86.75 %	86.18 %	86.28 %
2^{nd} Order	*87.33 %*	*87.28 %*	*87.74 %*	*87.11 %*	*87.04 %*
3^{rd} Order	83.62 %	85.36 %	85.51 %	85.65 %	84.76 %
4^{th} Order	84.41 %	85.69 %	85.71 %	85.98 %	85.45 %
Descriptor	Recognition Accuracy (%) Using KNN Classifier				
Order	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$
1^{st} Order (VDP)	85.08 %	85.45 %	85.96 %	86.03 %	85.88 %
2^{nd} Order	*85.34 %*	*85.75 %*	*86.62 %*	*86.44 %*	*85.95 %*
3^{rd} Order	82.23 %	84.47 %	84.94 %	84.86 %	84.52 %
4^{th} Order	82.13 %	84.48 %	84.64 %	84.71 %	84.42 %

4.1. YouTube Celebrities Dataset. YouTube Celebrities database is a large-scale video dataset which contains 1910 video sequences of 47 different celebrities (actors and politicians) that are collected from YouTube. The dataset is considered as one of the most challenging video databases due to the large illumination, pose, and expression variations as well as low resolution and motion blur. In this part, we evaluated the proposed HOVDP on all 47 celebrities, while some of the state-of-the-art compared methods were evaluated on some of the subjects (e.g., Yang et al. [23] use only the first 29 celebrities). Following the prior works [10, 22], for each subject three video sequences are randomly selected as the training data, with the other six video clips randomly selected for testing. While using the publicly available code of VLPQ technique [13], for each subject, six video sequences are randomly selected as the training data, with the other three video clips randomly selected for testing. For this publicly available code, we have used the default settings of all its parameters which yield the optimal performance. We conduct one experiment by random selection of training/testing data. The clips contain different numbers of frames (from 8 to 400) which have mostly low resolution and are highly compressed. Figure 4 shows some examples of cropped faces in this dataset.

Figure 4: YouTube Celebrities database. Each row represents different samples of one subject.

To show the effectiveness of the proposed HOVDP technique, we summarize the recognition rates via changing the neighborhood layers size (the order of the proposed approach) in the range $(1 - 4)$ where 1 means 3×3 window size, 2 means 5×5 window size, etc., varying the threshold t (the most prominent edge response values) in the range $\{t = 2, 3, \ldots, 6\}$, as well as comparing with VDP (the special case of HOVDP) in Table 1. From Table 1, it is found that $t = 4$ yields optimal performance for this dataset. Additionally, it is clear that the second order VDP improves the face recognition accuracy of the first order VDP in all test cases, while the third order and fourth order decrease the accuracy rates due to the fact that increasing the scale (the neighbors pixels) causes it to extract and fuse the information of different poses and different locations of the face components, which produces confused feature vectors. In addition, this increase of the descriptor order (the neighborhood layers size) would increase the feature vector length, which slows down the feature extraction processing speed.

The performance results of well-known face recognition algorithms like regularized nearest points (RNP) [23], sparse approximated nearest points between image sets (SANP) and its kernel extension (KSANP) [22], and projection metric learning on Grassmann manifold (PML) [10] combined with Grassmannian graph-embedding discriminant analysis (GGDA) [11] denoted as (PML-GGDA), with the proposed method HOVDP and the original VDP, as well as with the descriptor of the dynamic texture VLPQ [13] on this dataset, are presented in Table 2. Notice that the results we compared ours with are what we got from their original references, which are mentioned in the table. While for the VLPQ technique, the obtained results are based on its default settings, which yield the optimal performance. Meanwhile,

Table 2: Performance comparison of the proposed method with the algorithms on YouTube celebrities database.

Recognition Accuracy + Standard Deviation for Each Method					
SANP [22]	KSANP [22]	PML-GGDA [11]	VLPQ [13]	VDP [14]	HOVDP (Proposed)
55.64±5.7 %	65.4±5.5 %	70.32±3.6 %	82.97±0 %	86.75±0 %	*87.74±0 %*

Table 3: Recognition rates using only 50 frames of each Video as changing the threshold t and the proposed approach order on Honda/UCSD database.

Descriptor	Recognition Accuracy (%) Using libsvm Classifier				
Order	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$
1^{st} Order (VDP)	86.85 %	85.45 %	86.10 %	86.10 %	81.40 %
2^{nd} Order	*87.75 %*	*86.85 %*	*87.00 %*	*87.45 %*	*84.30 %*
3^{rd} Order	85.70 %	84.65 %	86.45 %	85.75 %	82.75 %
4^{th} Order	86.70 %	84.45 %	86.75 %	85.55 %	82.85 %
Descriptor	Recognition Accuracy (%) Using knn Classifier				
Order	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$
1^{st} Order (VDP)	88.75 %	90.11 %	89.65 %	88.65 %	88.55 %
2^{nd} Order	*89.25 %*	*90.30 %*	*90.10 %*	*89.65 %*	*88.80 %*
3^{rd} Order	86.25 %	87.05 %	88.70 %	87.60 %	87.85 %
4^{th} Order	87.00 %	88.05 %	88.40 %	87.55 %	87.50 %

Figure 5: Honda/UCSD dataset. Each row represents different samples of one subject.

a part of this dataset was used in RNP [23] and three video sequences were randomly selected as the training data, with the other three sequences randomly selected as the testing data.

4.2. Honda/UCSD Dataset. Honda/UCSD database consists of 59 video sequences of 20 different subjects. There are pose, illumination, and expression variations across the sequences for each subject. Each video consists of about 12 – 645 frames. Figure 5 has shown some examples [20]. Each row corresponds to an image set of a subject. In our experiment, we use the standard training/testing configuration provided in [19]. 20 sequences are used for training which means one video for each subject and the remaining 39 sequences for testing. We report results using all frames as well as with limited number of frames. Specifically, we conduct the experiments following the prior works [22, 23] by executing three parts of experiments, (1) using only the first 50 frames/video clip, (2) using only the first 100 frames/video clip, and (3) using all video frames. In case a set contains frames fewer than the selected ones, all frames are used for recognition process. The performance results of the proposed technique HOVDP as changing the number of neighborhood layers (the order of the proposed approach) in the range (1 – 4) along with changing the threshold t (the most prominent edge response values) in the range $\{t = 2, 3, \ldots, 6\}$ using set lengths 50 frames/clip, 100 frames/clip, and all frames/clip, respectively, are presented in Tables 3, 4, and 5.

The performance results of well-known video-based algorithms like SANP, KSANP, RNP, VLPQ, and the original VDP (the special case of the proposed technique) with the proposed method HOVDP on this dataset are presented in Table 6. Notice that the results we compared ours with are what we got from their original references, which are mentioned in the table. While for the VLPQ technique, the obtained results are based on its default settings, which provide the optimal performance. For our proposed methods, we select the value of t that yields optimal performance for the comparison.

5. Conclusion

In this paper, we introduced a new feature descriptor, namely, HOVDP. Throughout the performance evaluation in terms of face recognition accuracy, we found that HOVDP is

Table 4: Recognition rates using only 100 frames of each video as changing the threshold t and the proposed approach order on Honda/UCSD database.

Descriptor	Recognition Accuracy (%) Using libsvm Classifier				
Order	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$
1st Order (VDP)	87.86 %	87.09 %	87.99 %	87.99 %	84.96 %
2nd Order	*89.71 %*	*89.71 %*	90.21 %	88.89 %	86.25 %
3rd Order	87.89 %	87.83 %	90.86 %	88.99 %	89.34 %
4th Order	88.34 %	89.57 %	*91.68 %*	*89.36 %*	*90.20 %*
Descriptor	Recognition Accuracy (%) Using knn Classifier				
Order	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$
1st Order (VDP)	90.47 %	92.13 %	92.16 %	91.29 %	90.47 %
2nd Order	*91.79 %*	*92.56 %*	*93.08 %*	*92.58 %*	*91.66 %*
3rd Order	88.89 %	91.10 %	91.58 %	91.63 %	91.13 %
4th Order	88.99 %	91.50 %	91.84 %	91.74 %	91.61 %

Table 5: Recognition rates using all video frames as changing the threshold t and the proposed approach order on Honda/UCSD database.

Descriptor	Recognition Accuracy (%) Using libsvm Classifier				
Order	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$
1st Order (VDP)	94.75 %	95.09 %	95.96 %	95.82 %	94.05 %
2nd Order	*96.26 %*	*96.55 %*	*97.23 %*	96.57 %	95.74 %
3rd Order	95.08 %	94.42 %	96.23 %	96.55 %	95.32 %
4th Order	95.57 %	95.11 %	96.72 %	*97.13 %*	*95.84 %*
Descriptor	Recognition Accuracy (%) Using knn Classifier				
Order	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$
1st Order (VDP)	96.13 %	95.58 %	96.87 %	95.97 %	95.44 %
2nd Order	*96.52 %*	*96.87 %*	*97.08 %*	*96.61 %*	*96.23 %*
3rd Order	95.41 %	95.95 %	96.74 %	96.44 %	95.65 %
4th Order	95.42 %	96.11 }	96.63 %	96.51 %	95.86 %

Table 6: Performance comparison of the proposed method with the algorithms on Honda/UCSD database.

Number of Frames	Recognition Accuracy (%) for Each Method					
	SANP [22]	KSANP [22]	RNP [23]	VLPQ [13]	VDP [14]	HOVDP (Proposed)
50	84.62 %	87.18 %	87.18 %	55.00 %	90.11 %	*90.45 %*
100	92.31 %	94.87 %	94.87 %	70.00 %	92.27 %	*94.92 %*
All frames	*100 %*	*100 %*	*100 %*	80.00 %	96.87 %	97.29 %
Average	92.31 %	94.02 %	94.02 %	68.33 %	93.08 %	*94.22 %*

robust for video-based face recognition applications. With a video as input and a gallery of videos, we performed face recognition process throughout all the video clip frames. From the evaluation results, it has been found that the proposed HOVDP algorithm can successfully improve the accuracy rates compared to the original VDP in all test cases and exceed a set of state-of-the-art methods in most test cases. For the Honda/UCSD database, our proposed technique provides better recognition rates, although the other compared methods outperform ours in case all frames are used. Meanwhile, our proposed HOVDP beats the others in case smaller sets are used, which often occurs in real-world applications. For example, the tracking of a face may fail for a long video sequence when only the first part of the video sequence is available for the classification.

Data Availability

The video sequences that have been used in this paper may be found in the following links: YouTube Celebrities dataset at http://seqamlab.com/software-and-data/ Honda/UCSD dataset at http://vision.ucsd.edu/~iskwak/HondaUCSDVideoDatabase/HondaUCSD.html.

Disclosure

This manuscript was prepared based on the research that has been conducted as a part of the doctoral dissertation work of Almabrok Essa at the University of Dayton, Dayton, Ohio. The original dissertation document is available at https://etd.ohiolink.edu/!etd.send_file?accession=dayton150-0901918995427&disposition=inline.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] U. Raju, A. S. Kumar, B. Mahesh, and B. E. Reddy, "Texture classification with high order local pattern descriptor: local derivative pattern," *Global Journal of Computer Science and Technology*, vol. 10, no. 8, pp. 72–76, 2010.

[2] A. E. Essa and V. K. Asari, "Local directional pattern of phase congruency features for illumination invariant face recognition," in *Proceedings of the Optical Pattern Recognition XXV*, Baltimore, Md, USA, 2014.

[3] A. Essa and V. K. Asari, "Face recognition based on modular histogram of oriented directional features," in *Proceedings of the 2016 IEEE National Aerospace and Electronics Conference (NAECON) and Ohio Innovation Summit (OIS)*, pp. 49–53, Dayton, OH, USA, July 2016.

[4] A. Essa and V. Asari, "Local boosted features for illumination invariant face recognition," in *Proceedings of the International Conference on Electronic Imaging, Imaging and Multimedia Analytics in a Web and Mobile World*, pp. 70–73, Society for Imaging Science and Technology, 2017.

[5] J. Suneetha, "A survey on video-based face recognition approaches," *International Journal of Application or Innovation in Engineering And Management (IJAIEM)*, vol. 3, no. 2, pp. 208–215, 2014.

[6] G. Zhao and M. Pietikäinen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[7] S. Nagendra, R. Baskaran, and S. Abirami, "Video-based face recognition and face-tracking using sparse representation based categorization," *Procedia Computer Science*, vol. 54, pp. 746–755, 2015.

[8] M. Alajmi, K. Awedat, A. Essa, F. Alassery, and O. S. Faragallah, "Efficient face recognition using regularized adaptive non-local sparse coding," *IEEE Access*, vol. 7, pp. 10653–10662, 2019.

[9] J. Zheng, R. Ranjan, C.-H. Chen, J.-C. Chen, C. D. Castillo, and R. Chellappa, "An automatic system for unconstrained video-based face recognition," https://arxiv.org/abs/1812.04058, 2018.

[10] Z. Huang, R. Wang, S. Shan, and X. Chen, "Projection metric learning on grassmann manifold with application to video based face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pp. 140–149, USA, June 2015.

[11] M. T. Harandi, C. Sanderson, S. Shirazi, and B. C. Ovell, "Graph embedding discriminant analysis on Grassmannian manifolds for improved image set matching," in *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011*, pp. 2705–2712, USA, June 2011.

[12] R. Vemulapalli, J. K. Pillai, and R. Chellappa, "Kernel learning for extrinsic classification of manifold features," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2013*, pp. 1782–1789, June 2013.

[13] J. Päivärinta, E. Rahtu, and J. Heikkilä, "Volume local phase quantization for blur-insensitive dynamic texture classification," in *Image Analysis*, vol. 6688 of *Lecture Notes in Computer Science*, pp. 360–369, Springer, Heidelberg, Germany, 2011.

[14] A. E. Essa and V. K. Asari, "Video-to-video pose and expression invariant face recognition using volumetric directional pattern," in *Proceedings of the 10th International Conference on Computer Vision Theory and Applications, VISAPP 2015*, pp. 498–503, Berlin, Germany, March 2015.

[15] A. Essa, P. Sidike, and V. Asari, "Volumetric directional pattern for spatial feature extraction in hyperspectral imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 14, no. 7, pp. 1056–1060, 2017.

[16] P. Viola and M. J. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[17] R. A. Kirsch, "Computer determination of the constituent structure of biological images," *Computers and Biomedical Research*, vol. 4, no. 3, pp. 315–328, 1971.

[18] M. Kim, S. Kumar, V. Pavlovic, and H. Rowley, "Face tracking and recognition with visual constraints in real-world videos," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 1–8, USA, June 2008.

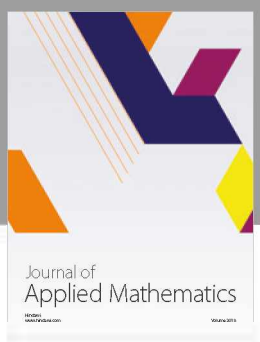
[19] K.-C. Lee, J. Ho, M.-H. Yang, and D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," in *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 313–320, USA, June 2003.

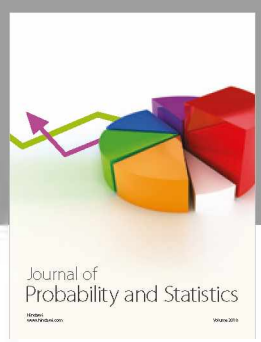
[20] K. Lee, J. Ho, M. Yang, and D. Kriegman, "Visual tracking and recognition using probabilistic appearance manifolds," *Computer Vision and Image Understanding*, vol. 99, no. 3, pp. 303–331, 2005.

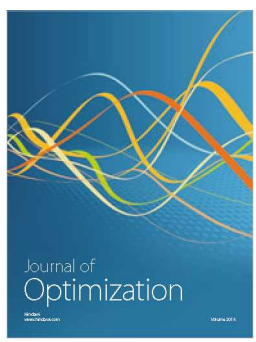
[21] C. Chang and C. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, article 27, 2011.

[22] Y. Hu, A. S. Mian, and R. Owens, "Face recognition using sparse approximated nearest points between image sets," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1992–2004, 2012.

[23] M. Yang, P. Zhu, L. Van Gool, and L. Zhang, "Face recognition based on regularized nearest points between image sets," in *Proceedings of the 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013*, pp. 1–7, China, April 2013.







Hindawi
Submit your manuscripts at
www.hindawi.com