Gait Recognition using Dynamic Texture Descriptors

Behnaz Abdolahi, Niloofar Gheissari
Dept. Electrical and Computer Engineering
Isfahan University of Technology
Isfahan, Iran
b.abdolahi@ec.iut.ac.ir, n.gheissari@cc.iut.ac.ir

***Abstract*— the human motion analysis is an attractive topic in biometric research. Common biometrics is usually time-consuming, limited and collaborative. These drawbacks pose major challenges to recognition process. Recent researches indicate people have considerable ability to recognize others by their natural walking. Therefore, gait recognition has obtained great tendency in biometric systems. Gait analysis is inconspicuous, needs no contact, cannot be hidden and is evaluated at distance.**

This paper presents a bag of word method for gait recognition based on dynamic textures. Dynamic textures combine appearance and motion information. Since human walking has statistical variations in both spatial and temporal space, it can be described with dynamic texture features. To obtain these features, we extract spatiotemporal interest points and describe them by a dynamic texture descriptor. To get more suitable results, we extend LBP-TOP as a rotation invariant dynamic texture descriptor. Afterwards, hierarchical K-means algorithm is employed to map features into visual words. At result, human walking represent as a histogram of video-words occurrences. We evaluate the performance of our method on two dataset: the KTH dataset and IXMAS multiview dataset.

Keywords-human motion analysis; gait recognition; spatiotemporal interest point; dynamic texture; visual dictionary; bag of words.

I. INTRODUCTION

The analysis of human walking is one of the recent attractive topics in computer vision. It has obtained great interest because of its applications in sport video analysis, robot guidance and gait recognition[1]. In this paper, we propose a method for gait recognition as a biometrics.

Common biometrics is usually time-consuming, limited and collaborative. Recent researches purpose a way to detect threats without attracts the attention of people especially in public places like airports, banks and subway stations. Psychological studies indicate that people have significant ability to recognize others by the way they walk. Therefore, gait recognition has obtained great tendency in biometric systems. Gait recognition is the task of identifying people based on a video sequence taken from their walking. Gait analysis is inconspicuous, needs no contact, cannot be hidden and is evaluated at distance.

Researchers have proposed many approaches for human walking analysis. Recent approaches have great interest to incorporate both motion and appearance information in spatiotemporal domain. This is expected to improve the recognition rate in compare to methods which rely on only motion or appearance information. In the appearance based methods, extracting an accurate silhouette demands a perfect segmentation. So these methods are sensitive to shape deformation due to different clothing or carrying accessories such as handbag or ball. Furthermore, these methods have difficulties in handling occlusion and camera calibration. On the other hand, motion based methods disregard human body shape. These methods are sensitive to image sequences changes in temporal domain such as motion speed changes. Using spatiotemporal volume is a common method that encodes motion and appearance information simultaneously. Bobick and Davis [2] employed Motion Energy Images (MEI) and Motion History Images (MHI) to recognize human motion. Weinland et al. in [3] generalized 2D motion templates into 3D and used motion history volume for human motion representation.

Dynamic textures are sequence of textures which have statistical variations in temporal domain. So dynamic texture descriptors naturally encode motion information. Applying them on textured regions, encodes appearance information as well. Therefore, dynamic texture descriptors can be used to describe the local features of human motion in both spatial and temporal domains. Local Binary Pattern from Three Orthogonal Planes (LBP –TOP) is a dynamic texture descriptor that utilizes spatiotemporal analysis for human walking description. Recently, LBP-TOP has been used in many applications such as facial expression[4], visual speech[5], action recognition[6] and gait recognition[1]. This dynamic texture descriptor is robust to rotation and scale changes.

Extracting as much discriminative features as possible is the main prerequisite to render the best description. Human motion features can be represented globally or locally. In global methods, the observation is described as a whole, so all image information has equal effect on the result. In this method, silhouette is extracted by background subtraction and then the entire interest region specifies the input of image descriptor. For example, blank and et al. in [7] tracked the human silhouette over time and constructed a 3D space-time volume by stacking silhouettes of image sequences. Then, they described this 3D volume globally as a 3D shape. Using all image information make global representation powerful, also cause complexity in calculations. Sensitivity on occlusion, noise and variations of view point are other disadvantages of this method.

In local methods, the observation is described as a collection of local features on image sequences. These features are positions where mutation occurs in spatiotemporal domain. The local interest regions around these features specify the input of image descriptor. Laptev and Lindeberg [8], extended Harris corners to spatiotemporal domain based on scale-space theory. They computed the spatiotemporal events by maximizing a normalized Laplacian operator over spatial and temporal domains. Local methods do not require background subtraction and make the recognition robust to partial occlusion or fragmented silhouette. Although local methods extract less structural information than global methods, they have the advantage of being compact and discriminative.

In this paper, we propose a bag of visual word method for analysis of human natural walking using dynamic texture descriptors. In our method, we use the advantages of local methods to extract the proper number of local human walking features. Then, we apply a dynamic texture descriptor to describe these features in spatiotemporal way. Afterwards, we use a clustering algorithm to construct the visual dictionary of visual words. In the bag of visual words model, local features are mapped into visual words. Therefore, a walking video sequence can be represented as a histogram of visual words frequencies.

The most computational cost of our method pertains to description step that depends on the number of extracted features. We propose a computationally simple descriptor that utilizes local binary patterns in three orthogonal planes to analyze human walking features.

In this paper, we explain our approach in section 2. In section 3, we illustrate and discuss our experimental results which confirm the success of the proposed method and finally conclusion is in section 4.

II. PROPOSED METHOD

Human motion analysis is a multistage process. A major step in this process is feature extraction. So our method relies on STIP features as a local method. Then, in description step, we extend LBP-TOP descriptor that is based on dynamic textures and represent human motion in spatiotemporal way. In the third step, the hierarchical K-means algorithm as a clustering algorithm will be applied to obtain the visual dictionary of video-words. At last, in pattern matching step, we use SVM classifier to compare feature vectors extracted from test sequence with training models.

A. Feature Extraction

An abstract representation of image patterns can be shown by local image features. These features or interest points represent mutation in spatial and temporal domains. Usually, static points and points with monotonous motion will not be extracted as interest points. Since our algorithm works on only interest points, there is no need to subtract background and this makes our method computationally simple and suitable for many applications.

In this paper, we utilize spatiotemporal corner detector that is performed by Laptev and Lindeberg [8]. They extend 2D Harris operator [9] to 3D interest point detector. They construct a 3D Harris matrix which includes temporal gradients and detects spatiotemporal corners as Spatial Temporal Interest Points (STIP) (see Figure 1). These are points at which significant change occurs in both space and time. This means features selected by STIP not only undergo an intensity change, but also they undergo a change in the magnitude of motion velocity or its direction. This results in detecting corners where a significant motion event has occurred.

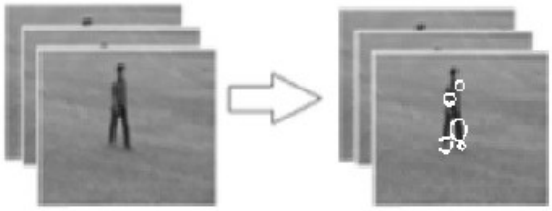
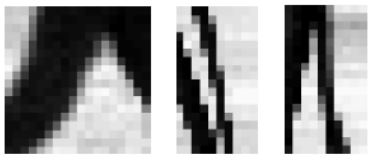


Figure 1.Interest point detection

B. Extended LBP-TOP Descriptor

LBP-TOP is an extension of basic LBP operator to describe dynamic textures. Since human walking is considered as a dynamic texture, it can be described with a dynamic texture descriptor. In this way three orthogonal planes with specific dimensions are assumed around each extracted interest point: XY, XT and YT (see Figure 2). XY plane is in spatial space and XT, YT planes are in temporal spaces.



(a) (b) (c)

Figure 2. Extracted planes: (a) XY, (b) XT, (c) YT [10].

For gait recognition using appearance is more informative than using motion. Then, we have given more weight to appearance (spatial space) than motion (temporal spaces). For this target, as is shown in Figure 3, we assumed three planes in spatial space (plane for points in ¼, ½ and ¾ of spatial space) and one plane for each temporal space. The dimensions of these planes are specified as 25*25 for each spatial plane (XY planes) and 25*20 for each temporal plane (XT and YT planes).

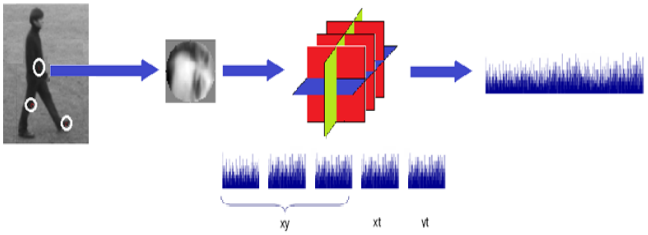


Figure 3. Orthogonal planes for interest regions

Then, a LBP based operator is applied on each plane and one histogram of binary codes is obtained for each one. After normalizing these histograms, the threefold weight is given to spatial central histogram and the equal weight is given to other spatial and temporal histograms. Therefore, the weight ratio of spatial space to temporal space will be 5 to 2. Finally, each interest point is described by concatenating these histograms.

LBP -TOP descriptor is rotation and scale invariant but is sensitive to global representation. Heretofore, LBP-TOP is used in some approaches[11] that they apply normal LBP operator for describing every planes. In this paper, we use Local Binary Pattern Histogram Fourier operator (LBP-HF) to describe spatial planes and rotation invariant LBP operator (LBP^{riu2}) to describe temporal planes.

1) *LBP operator*

Local Binary Pattern (LBP) is a descriptor for static textures[12, 13]. LBP operator provides a binary code which is the result of comparing a neighborhood of pixels with central pixel. In the basic LBP operator, as is shown in Figure 4-a, a neighborhood of pixels will be assumed around each pixel. The different size of circular neighborhoods with variety in radius and number of points is indicated in Figure 4-b.

Then, the gray value of center pixel (g_c) is subtracted from gray value of neighborhood points (g_p). The value of neighborhood points will be one for positive and zero for negative results. After assigning 2^P factor to neighbors, the LBP binary code of central pixel is computed as:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s\left(g_p - g_c\right) 2^p,\ s(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \qquad (1)$$

As a result, the texture image is represented by a histogram (k is the maximum value of LBP pattern):

$$H(k) = \sum_{i=1}^{N} \sum_{j=1}^{M} f\left(LBP_{P,R}(i,j), k\right), k \in [0, K] \qquad (2)$$

$$f(x,y) = \begin{cases} 1, & x = y \\ 0, & otherwise \end{cases}$$

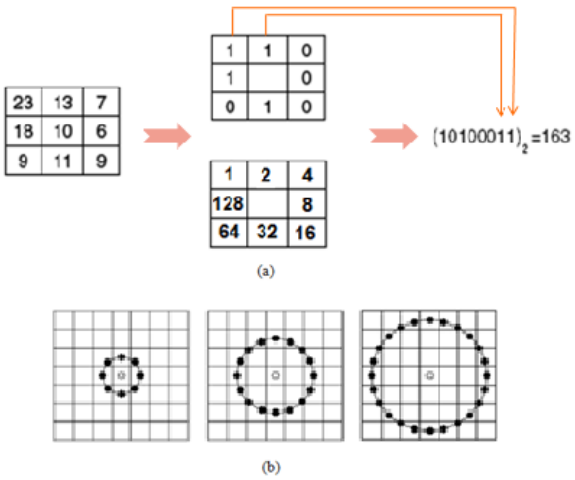


Figure 4.(a) Basic LBP (b) Three circular neighborhoods: (8,1), (16,2), (24,3)

The uniform LBP patterns are fundamental patterns of image texture (U<=2) (see Figure 5). The U value of LBP extension refers to the number of spatial transitions of binary codes in circular neighborhood (bitwise 0/1 changes):

$$U(LBP_{P,R}) = |s(g_{P-1} - g_c) - s(g_0 - g_c)| + \sum_{p=1}^{P-1} |s(g_{P-1} - g_c) - s(g_0 - g_c)| \qquad (3)$$

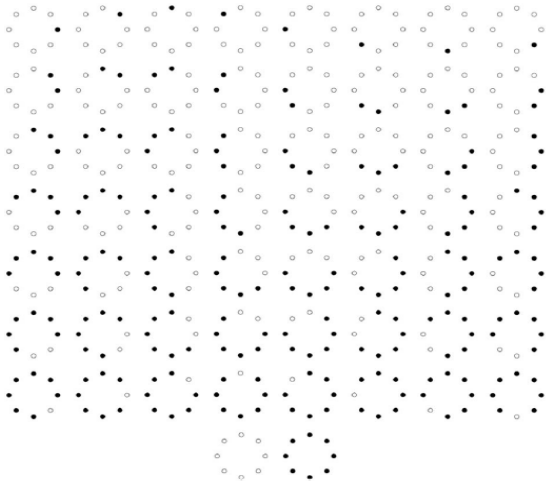


Figure 5.Uniform LBP Patterns (P=8) [14]

By rotating an image, the gray value of neighbors revolves along the perimeter of circular neighborhood of central pixel. To eliminate the effect of rotation and assimilate all rotated LBP patterns, we define:

$$LBP_{P,R}^{ri} = min\{ROR\left(LBP_{P,R}, i\right) \mid i = 0,1,,\ldots,P-1\} \qquad (4)$$

As is shown in Figure 5, the patterns in the same row are rotated version of each other, so they are the same. Therefore, the rotation invariant LBP extension is defined as:

$$LBP_{P,R}^{riu2} = \begin{cases} \sum_{p=0}^{P-1} s\left(g_p - g_c\right) & if\ \ U\left(LBP_{P,R}\right) \leq 2 \\ P+1 & otherwise \end{cases} \qquad (5)$$

2) LBP-HF operator

Local Binary Pattern Histogram Fourier (LBP-HF) as a globally rotation invariant descriptor is based on uniform local binary patterns. LBP-HF features have been calculated from discrete Fourier transforms of LBP histograms[15]. For this aim, uniform LBP histogram will be calculate over the whole region and then globally rotation invariant features will be obtain from discrete Fourier transforms of this histogram. Therefore, these features are invariant to rotations of the whole input but they still retain local rotation information. This information is about relative distribution of different orientations of uniform patterns.

Let $h_I(U_P(n,r))$ be consider as the uniform LBP histogram. By the rotation of the input image with $\propto$=a360/P, a=0,1,…,P-1 degree, the histogram will have a cyclic shift.

$$h_{I^{\alpha}}(U_P(n,r+a)) = h_I \qquad (6)$$

Let H(n,.) be the discrete Fourier transform of nth bin of the histogram:

$$H(n,u) = \sum_{r=0}^{P-1} h_I(U_P(n,r)) e^{-i2\pi ur/P} \qquad (7)$$

So Discrete Fourier coefficients will have a phase shift if the histograms cyclically shift.

This means, if $h'(U_P(n,r)) = h(U_P(n,r-a))$ then:

$$H'(n,u) = H(n,u)\, e^{-i2\pi ua/P} \qquad (8)$$

As $\overline{H(n,u)}$ denotes the complex conjugate of $H(n,u)$, for $1 \leq n_1$ and $n_2 \leq P-1$, we will have:

$$H'(n_1,u)\overline{H'(n_2,u)} = H(n_1,u)e^{-i2\pi ua/P}\overline{H(n_2,u)e^{-i2\pi ua/P}} = H(n_1,u)\overline{H(n_2,u)} \qquad (9)$$

Therefore, LBP-HF features:

$$LBP^{u2} - HF(n_1,n_2,u) = H(n_1,u)\overline{H(n_2,u)} \qquad (10)$$

are invariant to cyclic shifts of uniform LBP histogram and so, they are invariant to globally rotation of the input image.

3) Multiresolution Analysis

To increase the performance of our method, we also use multiresolution analysis. This method combines multiple kernels of LBP based operators with different P and R parameters. In this way, as is shown in Figure 6, the final histogram is gotten by concatenating the derived histogram of each kernel. In this paper, P_x= P_y= P_t=8 , R_x= R_y= R_t=1 are proposed for the first kernel and P_x= P_y= P_t=8 , R_x= R_y= R_t=2 for another.

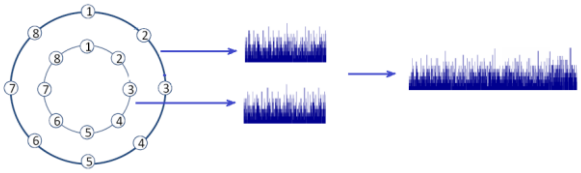


Figure 6. multiresolution analysis

C. Creating Visual Dictionary

After description step, each interest point is represented as a histogram. Since each image sequence has many extracted local interest points, so each image sequence have been exposed as a collection of histograms. It may cause some problems to compare sequences. Creating a visual dictionary can be use to overcome this problem.

The concept of visual dictionary is borrowed from image segmentation and retrieval. Visual dictionary has usually been initialized by applying an unsupervised clustering algorithm. hierarchical K-means is a prevalent clustering algorithm that can be used to assign the local features into visual words [16] (see Figure 7).

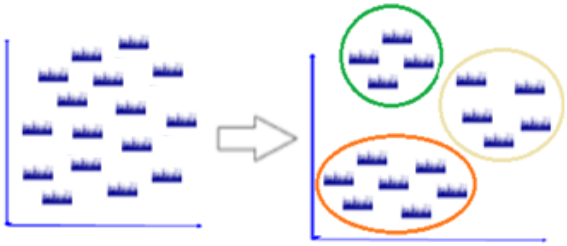


Figure 7. Creating visual dictionary

After creation of visual dictionary, each image sequence will be represented by a histogram of visual words occurrences as a feature vector.

1) Hierarchical K-means algorithm

Applying hierarchical K-means algorithm constructs a K-d tree. Unlike the original K-means algorithm that K specifies the total number of clusters, in this algorithm K defines the branch factor of the d levels tree. In this method, original K-means algorithm is recursively applied to collection of histograms as the data. In each run, K cluster centers are obtained and then data is divided into K groups based on nearest distance with cluster centers. Next, the original K-means algorithm is applied to each group as well. At last, the K^d number of clusters is obtained.

2) Stop list analogy

Using stop list analogy, help us to have more desirable visual dictionary and improve the performance of our method. In this analogy, visual words are composed according to words frequency. Most frequent or infrequent visual words can cause mismatch. Then, by using the stop list, these visual words are removed. In our method, visual words in the same cluster have

great dependency. By contrast, visual words with uniform distribution are less informative and cause mismatch. Then, they will be omitted. Also the size of visual dictionary can be reduced by merging similar visual words to gain a balanced between intra-class compactness and extra-class severability.

D. Classification

In this paper, we use support vector machine (SVM) classifier that is among the most popular and powerful classification algorithms in computer vision.

SVM classifier is based on maximizing margin between classes. For this target, SVM classifier supposes a hyper-plane for two-class problems or set of hyper-planes for multiclass problems. The hyper-plane has maximum distance to nearest object of each class. Nonlinear SVM classifiers use the kernel functions to map objects to higher dimensional space. Then, the hyper-plane is supposed in the transformed space. In this paper, we use polynomial kernel in SVM classification algorithm.

III. Experimental Results

We evaluate the performance of our method on two dataset: the KTH dataset[11] and IXMAS multiview dataset[17].

These are action datasets and we are the first to test gait recognition method on these datasets. These datasets have less limitation in walking and actions are done in different condition of view point, scale and light. We use these datasets to estimate the efficiency of our method on different condition. Whereas, our method is robust to rotation, view point, scale and appearance changes, we have gotten a reasonable performance on these datasets.

A. KTH dataset

KTH is one of the most popular datasets has been used for human movement analysis. In this dataset 25 people perform 6 actions. These actions have been performed under four different conditions: outdoors with static camera (normal), outdoors with a camera zooming in and out (scale variation), outdoors with different clothing (appearance changes) and indoors (light variation). The background is static as well.

We first experimented [18] with KTH dataset. In Our experiment, we use walking video sequences in KTH dataset under three conditions of normal, scale variation and clothing changes (see Figure 8). So the indoor video sequences are excluded.

Figure 8. Some samples of walking action in KTH dataset

For KTH dataset, we use K=2 and d=10 parameters and the size of visual dictionary will be 1024 visual words. Then, by applying stop list analogy, the size of visual dictionary will be 700 visual words.

We test our method on gait recognition in two ways. TABLE 1 reports the results of first test that learning is performed on video sequences under on condition and testing is performed under one another condition. In this case, the average rate is 70.2 %. On second test, we implement learning on video sequences under two conditions and testing under other condition. The results are reported in TABLE 2 and average rate is 77.3 %. In latter experiment, due to more training information, average efficiency is better than former experiment.

TABLE 1. Results reported on gait recognition for KTH database for first test

Train	Normal	Normal	Scale variation	Scale variation	Clothing changes	Clothing changes
Test	Scale variation	Clothing changes	Normal	Clothing changes	Normal	Scale variation
Results	**72.8%**	**77.3%**	**71.5%**	**62%**	**74%**	**63.7%**

TABLE 2. Results reported on gait recognition for KTH database for second test

Train	Scale variation And Clothing changes	Normal and Clothing changes	Normal and Scale variation
Test	Normal	Scale variation	Clothing changes
Results	**81.7%**	**73%**	**77.2%**

The results in TABLE 1 indicate normal video sequence contain information from other video sequences. It is similar to scale variation video sequence in appearance but is different in scale and view point. On the other hand, it is similar to appearance variation video sequence in scale and view point but is different on clothing. Using scale variation video sequence for train and clothing changes video sequence for test or contrariwise, due to difference on scale, view point and appearance, obtain the lowest results.

The results in TABLE 2 demonstrate that utilizing the more training information obtain better results. The best result will be obtained if we use clothing changes and scale variation video sequences for train. In this case, training process contain information of different scale, view point and clothing. However, using normal and appearance variation video sequences for train miss scale and view point information. Also by using normal and scale variation video sequences for train, clothing information will be omitted.

B. *IXMAS maltiview dataset*

IXMAS is a desired multiview dataset for human motion analysis. In this dataset 12 people perform 14 actions for three times. We select walking of 10 people for our experiment. These actions have been performed under five views that we use four views. Top view is not informative and is excluded (see Figure 9).

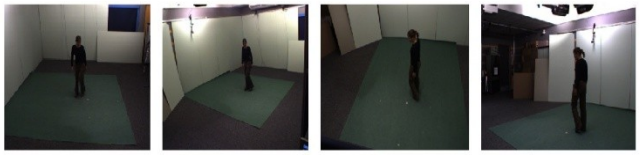


Figure 9. Some samples of walking action in IXMAS dataset in four views (from left to right be view 1 to 4).

As is shown in Figure 10, in this dataset each person walks circularly.

Figure 10. circular walking under view1

For IXMAS multiview dataset, we use K=2 and d=11 parameters and the size of visual dictionary will be 2048 visual words. Then, by applying stop list analogy, the size of visual dictionary will be 1700 visual words.

In IXMAS dataset, each person performs walking for three times in the same condition. Then, for each person, we will have three video sequences under each view. In our experiment, two video sequences is used for learn and other video sequence for test. The results are reported in TABLE 3 and average rate is 72.5 %.

TABLE 3. RESULTS REPORTED FOR IXMAS MULTIVIEW DATASET

	View1	View2	View3	View4
Results	**65%**	**70%**	**80%**	**75%**

Since messy scenes disturb feature extraction and obtain unsuitable features, sparse scenes get better results. Therefore, as is shown in TABLE 3, view3 is more instructive and get more acceptable result than others.

IV. CONCLUSION

In this paper, we proposed a bag of word approach for gait recognition using dynamic textures. Human walking has variations in spatial and temporal spaces so can be described by a dynamic texture descriptor. In this paper we extend LBP-TOP as a rotation invariant descriptor of features in local representation. Then, by using hierarchical K-means algorithm, the walking video sequence will be represented as a collection of video-words.

The performance of our method is studied on two public datasets: KTH and IXMAS multiview datasets and admirable results are obtained on both dataset. Since we are the first to test gait recognition method on these datasets, we could not compare our results with result of other methods.

REFERENCES

[1] M. Tistarelli, M. Nixon, V. Kellokumpu, G. Zhao, S. Li, and M. Pietikäinen, "Dynamic Texture Based Gait Recognition," in *Advances in Biometrics*. vol. 5558: Springer Berlin / Heidelberg,, pp. 1000-1009, 2009.

[2] J. W. D. Aaron F. Bobick, "The Recognition of Human Movement Using Temporal Templates," *IEEE transactions on pattern analysis and machine intelligence,* vol. 23, pp. 257-267, 2001.

[3] R. R. Daniel Weinland, Edmond Boyer, "Free Viewpoint Action Recognition using Motion History Volumes," *Elsevier Science,* 2006.

[4] Z. Guoying, "Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 29, pp. 915-928, 2007.

[5] G. Zhao, M. Pietikainen, and A. Hadid, "Local Spatiotemporal Descriptors for Visual Recognition of Spoken Phrases," *Proc. HCM 2007,* pp. 57-65, 2007.

[6] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing* pp. 976-990, 2010.

[7] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," 2005.

[8] I. Laptev and T. Lindeberg, "Space-time interest points," *International Conference on Computer Vision,* pp. 432-439, 2003.

[9] C. Harris and M. Stephens, "A combined corner and edge detector," in *Alvey vision conference,* p. 50,1988.

[10] R. Mattivi and L. Shao, "Human Action Recognition Using LBP-TOP as Sparse Spatio-Temporal Feature Descriptor," *CAIP 2009,* pp. 740–747, 2009.

[11] V. Kellokumpu, G. Zhao, and M. Pietikainen, "Dynamic textures for human movement recognition,", pp. 470-476, 2010.

[12] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence,* pp. 971-987, 2002.

[13] G. Zhao and M. Pietikainen, "Local binary pattern descriptors for dynamic texture recognition," *Pattern Recognition,* vol. 2, pp. 211-214, 2006.

[14] Z. Guo, L. Zhang, and D. Zhang, "Rotation invariant texture classification using LBP variance (LBPV) with global matching," *Pattern Recognition,* vol. 43, pp. 706-719, 2009.

[15] T. Ahonen, J. Matas, C. He, and M. Pietikainen, "Rotation invariant image description with local binary pattern histogram fourier features," *Image Analysis,* pp. 61-70, 2009.

[16] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* pp. 2161-2168, *2006*.

[17] J. Liu and M. Shah, "Learning human actions via information maximization,",CVPR 2008, pp. 1-8, 2008.

[18] B. Abdolahi, S. Ghasemi, and N. Gheissari, "Human Motion Analysis using Dynamic Textures," *Artificial Intelligence and Signal Proccessing, AISP*, 2012.