

Action Classification Using Weighted Directional Wavelet LBP Histograms

Maryam N. Al-Berry, Mohammed A.-M. Salem, Hala M. Ebeid,
Ashraf S. Hussein and Mohamed F. Tolba

Abstract The wavelet transform is one of the widely used transforms that proved to be very powerful in many applications, as it has strong localization ability in both frequency and space. In this paper, the 3D Stationary Wavelet Transform (SWT) is combined with a Local Binary Pattern (LBP) histogram to represent and describe the human actions in video sequences. A global representation is obtained and described using Hu invariant moments and a weighted LBP histogram is presented to describe the local structures in the wavelet representation. The directional and multi-scale information encoded in the wavelet coefficients is utilized to obtain a robust description that combine global and local descriptions in a unified feature vector. This unified vector is used to train a standard classifier. The performance of the proposed descriptor is verified using the KTH dataset and achieved high accuracy compared to existing state-of-the-art methods.

Keywords 3D stationary wavelet • Action recognition • Local binary pattern • Hu invariant moments • Spatio-temporal space Multi-resolution analysis

M.N. Al-Berry (✉) • M.A.-M. Salem • H.M. Ebeid • M.F. Tolba
Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt
e-mail: maryam_nabil@cis.asu.edu.eg

M.A.-M. Salem
e-mail: salem@cis.asu.edu.eg

H.M. Ebeid
e-mail: hlam@cis.asu.edu.eg

M.F. Tolba
e-mail: fahmytolba@gmail.com

A.S. Hussein
Faculty of Computing Studies, Arab Open University-Headquarters, 13033,
Kuwait, Kuwait
e-mail: ashrafh@acm.org

1 Introduction

Wavelets [1] and multi-scale techniques have been widely used in many image processing and computer vision applications [2, 3]. The most important property of wavelet analysis is the ability to analyze functions at different scales. Wavelets also provide a basis for describing signals accurately [4]. These properties of wavelet analysis have led to very successful applications within the fields of signal processing, image processing and computer vision. One of the fields that benefit from wavelets is the field of human action recognition [3, 5, 6].

In this way, wavelets have been used for spatio-temporal representation and description of human actions. In spatio-temporal methods, the action is represented by fusing spatial and temporal information into a single template. Spatio-temporal techniques can be divided into global and local techniques [7]. The advantages of both global and local representation techniques can be combined to obtain robust representations for human actions [8]. The 3D Wavelet Transforms [2, 6, 9] have been proposed and used to process video sequences as a 3D volume having time as the third dimension. The wavelet coefficients are then used to obtain global [3, 5, 10] or local [6] representations for the performed actions.

In this paper, a new method for action representation and description is proposed based on the wavelet analysis. The proposed method aims at extracting discriminative features and increasing the classification accuracy. The power of the 3D Stationary Wavelet Transform (SWT) is used to highlight spatio-temporal variations in the video volume and provides a multi-scale directional representation for human actions. These representations are described using a new proposed descriptor based on the concept of the Local Binary Pattern (LBP) [11], which is called the weighted directional wavelet LBP histogram. The proposed local descriptor is combined with Hu invariant moments [12] that are extracted from the Multi-scale Motion Energy Images (*MMEI*) [5] to get benefit from the global information contained in the wavelet representation. The performance of the proposed method is verified through several experiments using KTH dataset [13]. The rest of the paper is organized as follows: Sect. 2 reviews the background of the field and the state-of-the-art methods. The proposed representation and description method is illustrated in Sect. 3. In Sect. 4, the proposed method is applied, and the performance evaluation and comparison is performed. Finally, the work is concluded, and possible directions for improvement are pointed out in Sect. 5.

2 Related Work

The wavelet transforms [1, 14, 15] are being used in different fields such as signal, image and video compression as they provide the basis to describe signal features easily. The Decimated bi-orthogonal Wavelet Transform (DWT) [14] is probably the most widely used wavelet transform as it exhibits better properties than the

Continuous Wavelet Transform (CWT). It has much less scales than the CWT as only dyadic scales are considered. DWT is also very computationally efficient, since it requires $O(N)$ operations for N samples of data and can be easily extended to any dimension by separable products of the scaling and wavelet functions [16].

When dealing with images in the spatial domain, two-dimensional DWT is used to decompose an image into four sub-images; an approximation and three details. The approximation A is obtained by applying the low-pass filter to the two spatial dimensions of the image and represents the image at a coarse resolution. The three detailed images are the horizontal detail H , the vertical detail V and the diagonal detail D [2]. Wavelet analysis has been used in the context of 2D spatio-temporal action recognition. For example, Siddiqi et al. [17] have used 2D DWT to extract features from the video sequence. The most important features were then selected using a Step Wise Linear Discriminant Analysis (SWLDA). Sharma et al. [18] have represented short actions using Motion History Images (MHI) [19]. They used two dimensional, 3 level dyadic wavelet transforms and modified their proposed representation to be invariant to translation and scale, but they concluded that the directional sub-bands alone were not efficient for action classification. In [20], Sharma and Kumar have described the histograms of MHIs by orthogonal Legendre moments and modeled the wavelet sub-bands by Generalized Gaussian Density (GGD). They have found that the directional information encapsulated in the wavelet sub-band enhances the classification accuracy.

Wavelet transforms have also 3D extension. Rapantzikos et al. have used the 3D wavelet transform to keep the computational complexity low while representing dynamic events. They have used wavelet coefficients to extract salient features. They treated the video sequence as a spatio-temporal volume, and local saliency measures are generated for each visual unit (voxel). Shao et al. [21] have applied a transform based technique to extract discriminative features from video sequences. They have shown that the wavelet transform gave promising results on action recognition.

While the decimated wavelet transform has been successful in many fields, it has the disadvantage of translation variance, so to overcome this drawback and maintain shift invariance, the Stationary Wavelet Transform (SWT) was developed. It has been proposed in the literature under different names such as; un-decimated wavelet transform [4], redundant wavelet [22] and shift-invariant wavelet transform [23]. The SWT gives a better approximation than the DWT since it is linear, redundant and shift-invariant [2]. In, the 3D SWT has been proposed and used for spatio-temporal motion detection. A 3D SWT is applied to each set of frames, and then the analysis is performed along the x-direction, the y-direction and the t-dimension of the time varying data [2], which is formed using the input frames.

Based on this 3D SWT, the authors of [10] have proposed two spatio-temporal human action representations. The proposed representations were combined with Hu invariant moments as features for action classification. Results obtained using the public dataset were promising and provide a good step towards better enhancements. They have proposed a directional global wavelet-based

representation of natural human actions [3] that utilizes the 3D SWT [2] to encode the directional spatio-temporal characteristics of the motion.

Wavelets have also been used in combination with local descriptors such as Local Binary Patterns (LBPs) [11, 24] to describe texture images. LBPs are texture descriptors that have proved to be robust and computationally efficient in many applications [21, 25, 26]. The most important characteristic of LBPs is that it is not only computationally efficient, but also invariant to gray level changes caused by illumination variations [11]. The original LBP operator labels the pixels of an image by thresholding the 3×3 neighborhood of each pixel with the center value and considering the result as a binary number. The LBP was extended to operate on a circular neighborhood set of radius R . This produces 2^P different binary patterns for the P pixels constituting the neighborhood. Uniform LBP was defined as the pattern that contains at most 2 transitions for 1 to 0 or from 0 to 1. These patterns describe the essential patterns that can be found in texture. The disadvantage of LBP is lack of directional information.

In this paper, the 3D SWT is combined with LBPs and employed in the field of human action recognition. A new descriptor is proposed for describing human actions represented by the wavelet coefficients. The new descriptor combines the directional information contained in the wavelet coefficients in a weighted manner using the entropy value. The new local descriptor is expected to provide discriminative local features for the human actions.

3 Proposed Action Representation and Description

In this section, the proposed approach for extracting features is described. The proposed approach uses the power of the 3D SWT to detect multi-scale directional spatio-temporal changes and describes these changes using a weighted Local Binary Pattern (LBP) histogram. The weighted directional wavelet LBP is fused with moments to maintain global relationships. The proposed method is illustrated in Fig. 1.

3.1 *Weighted Directional Wavelet Local Binary Pattern Histogram*

The proposed descriptor is obtained by treating the video sequence as a 3D volume; considering time the third dimension. First, the video sequence is divided into disjoint blocks of frames and these blocks are supplied to the 3D SWT. The number of frames in one block depends on the intended number of wavelet analysis levels (resolutions). The number of frames is obtained by $L = 2^J$, where J is the number of resolutions. The 3D SWT produces the spatio-temporal detail coefficients

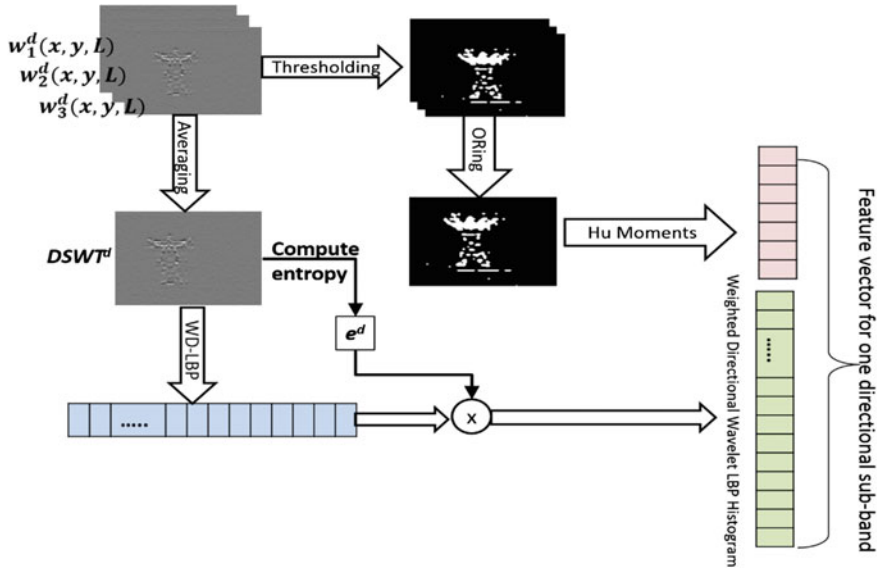


Fig. 1 Illustration of the proposed method

$w_j^d(x, y, t)$, where (x, y) are the spatial coordinates of the frames, t is the time, $j = 1, 2, \dots, J$ is the resolution (scale) and d is the sub-band orientation, and the approximation coefficient $c_j(x, y, t)$ computed by the associated scaling function as described in [2].

The detail sub-bands $w_j^d(x, y, t)$, $d = (5, 6, 7)$ are used for representing the action as they contain highlighted motion detected in the temporal changes that happened along the t -axis. Sub-band 4 also highlights temporal changes along with global spatial approximation. It is not used here because the focus is on investigating the effect of using local details only. The resulting coefficients are used to obtain a directional multi-scale stationary wavelet representation for the action by averaging the wavelet coefficients of the obtained resolutions at time $t = L$. This results in three Directional multi-scale Stationary Wavelet Templates $DSWT^d$, $d = (5, 6, 7)$, for the three detail sub-bands.

$$DSWT^d(x, y) = \frac{\sum_{j=1}^3 w_j^d(x, y, L)}{J}, d = (5, 6, 7) \quad (1)$$

Each directional stationary wavelet template is treated as texture images and the Local Binary Pattern is used to extract local features from it. The Directional Wavelet—LBP (DW-LBP) is computed as follows:

$$DW - LBP_{(P,R)}^d = \sum_{p=0}^{P-1} S(f_p^d - f_c^d) 2^p \quad (2)$$

where (P, R) denotes a neighborhood of P equally spaced sampling points on a circle of radius R , $S(z)$ is the thresholding function

$$S(z) = \begin{cases} 1, & z \geq 0 \\ 0, & z < 0 \end{cases} \quad (3)$$

f_c^d is the center pixel in the defined neighborhood in sub-band d , f_p^d , $p=0, 1, \dots, P-1$ are the pixels in the neighborhood. The normalized histogram of the obtained Directional Wavelet—Local Binary Pattern ($DW - LBP$) is computed and the entropy value [27] for each directional wavelet template (e^d) is computed and is used to give a weight for the computed histogram by multiplying it by the histogram values.

3.2 Global Description Using Invariant Moments

To obtain global description, the coefficients of the three selected sub-bands are thresholded to produce motion energy images. The motion images obtained at time $t=L$ of the J different scales are fused into three directional multi-scale motion energy images ($MMEI^d(x, y, t)$) in the same way proposed in [3]. These sub-band motion energy images encode the directional motion energy during the processed L frames. These coefficients of the different scales are threshold to obtain motion images $O_j^d(x, y, t)$

$$O_j^d(x, y, t) = \begin{cases} 1 & \text{if } w_j^d(x, y, t) > \tau \text{ for some } j, d \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

The threshold value τ is computed automatically using Otsu's thresholding technique [27]. This technique is completely based on the computation of the image histogram and maximizes the inter-class variance [27]. By fusing the motion images obtained from the different resolutions for each sub-band, multi-scale motion energy images ($MMEI^d(x, y, t)$) is obtained. These motion energy images are described using the seven Hu moments [12]. The Hu moments are known to be good global shape descriptors and are translation, scale, mirroring and rotation invariant [27].

To compute the set of invariant moments, the central moment of order $p+q$ (μ_{pq}) is computed by:

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q MMEI^d(x, y) \quad (5)$$

where, $\bar{x} = \frac{M_{10}}{M_{00}}$, $\bar{y} = \frac{M_{01}}{M_{00}}$, and M_{ij} are the raw image moments calculated by:

$$M_{ij} = \sum_x \sum_y x^i y^j MMEI^d(x, y) \quad (6)$$

The set of seven invariant moments are then computed as described in [27] and constitutes the second part of the final feature vector.

4 Experimental Results

This section describes the performed experiments and demonstrates the achieved results. First, the used datasets are described, then the experimental results are presented.

4.1 Human Action Dataset

The performance evaluation of the proposed method is carried out using the KTH dataset [13]. This dataset contains samples for six actions: “boxing”, “hand-clapping”, “hand-waving”, “jogging”, “running” and “walking”. These actions are performed by 25 different subjects in four different scenarios. The first scenario s1 is “outdoors”, s2 is “outdoors with scale variation”, s3 is “outdoors with different clothes” and s4 is “indoors”. Samples of the “boxing” action in the four different scenarios are shown in Fig. 2.

4.2 Experiments and Results

A simple K-Nearest Neighbor (KNN) classifier with ($K=1$) was used for testing the accuracy of the proposed features. The used KNN classifier uses Euclidean

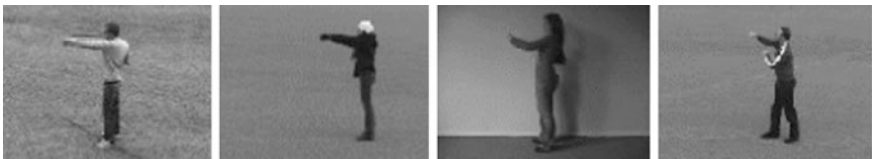


Fig. 2 Samples of the “boxing” action in the four different scenarios

distance. Leave-one-out cross-validation is used to calculate the classification accuracy. The descriptor is computed in (8,5) neighborhood, i.e., the radius of the circular neighborhood is 5 and 8 sampling points are used on the circle. These values for the radius and the sampling points were chosen empirically. The number of bins in the weighted histogram of DW-LBP is 59, as only uniform patterns are considered. The final feature vector length is 66 for each sub-band after combination with the seven Hu invariant moments.

The four different scenarios were first examined separately. In another experiment, the overall performance is measured by using the samples of the four scenarios. Two classification schemes were used in the evaluation. First, the features of the three directional sub-bands are combined in a single feature vector and this feature vector is used to train the classifier. In the other scheme, three classifiers are trained; each one using one feature vector of the three sub-bands and the class assigned to the test pattern is selected by majority voting between the three classifiers.

Table 1 lists the classification accuracy obtained for the four scenarios and the overall accuracy using the two classification schemes. In the case of the combined feature vector (CFV), scale variations and variations in clothes didn't affect the classification accuracy. Using voting the second scenario recorded less accuracy than the combined feature vector. It can be seen that the proposed method recorded a high classification accuracy consistently using the two classification schemes.

Another experiment was carried out using the gait actions only. The proposed method achieved a high accuracy of 97.33 % using the combined feature vector while gait actions are 100 % accurately classified using the voting scheme. The results are shown in Table 2.

Table 3 shows a comparison between the proposed method and existing state-of-the-art methods. The proposed method outperformed existing techniques, achieving higher classification accuracy. All compared techniques used leave-one-out cross validation.

Table 1 Accuracy using combined feature vector versus voting between directional feature vectors

Scenario	Combined feature vector %	Voting %
s1	94.67	95.33
s2	94.67	92.00
s3	94.67	94.67
s4	96.00	96.00
Overall	95.00	94.50

Table 2 Accuracy using combined feature vector versus voting for gait actions

Scenario	Combined feature vector %	Voting %
s1	97.33	100
s2	96.00	100
s3	98.67	100
s4	97.33	100
Overall	96.00	100

Table 3 Performance comparison with existing techniques

Method	Accuracy %
Proposed method (CFV)	95.00
Proposed method (voting)	94.50
Kong et al. (2011) [28]	88.81
Bregonzio (2012) [29]	94.33
Gupta et al. (2013) [30] (gait actions Only)	95.10
Proposed method (gait actions only)(CFV)	96.00
Proposed method (gait actions only)(voting)	100

5 Conclusions

In this paper, a new human action descriptor is proposed. The proposed descriptor combines the power of wavelets in highlighting directional variations and the simplicity and robustness of local binary patterns in describing local structures. The new local descriptor is based on the 3D stationary wavelet transform that highlights spatio-temporal variations representing human actions in video sequences. A weighted combination of directional features extracted from the wavelet coefficients is fused with global moments and used for describing actions. The proposed features are tested on a public dataset, and their accuracy is verified using a standard classifier. The proposed method outperformed state-of-the-art methods achieving 95 % average classification accuracy for all actions using combined feature vector and an average accuracy of 100 % for gait actions using voting.

Future work may include block -based processing of the wavelet multi-scale templates and extending the new descriptor to the 3D domain.

References

1. Ahad, M., Tan, J., Kim, H., Ishikawa, S.: Motion history image: its variants and applications. *Mach. Vis. Appl.* **23**, 255–281 (2012)
2. Al-Berry, M.N., Ebied, H.M., Hussein, A.S., Tolba, M.F.: Human action recognition via multi-scale 3D stationary wavelet analysis. In: *Hybrid Intelligent Systems 2014 (HIS'14)*. Kuwait (2014)
3. Al-berry, M.N., Salem, M.A.-M., Ebeid, H.M., Hussein, A.S., Tolba, M.F.: Action recognition using stationary wavelet-based motion images. In: *IEEE conference on Intelligent systems 14*, pp. 743–753. Warasaw, Poland: Springer International Publishing (2014)
4. Al-Berry, M.N., Salem, M.A.-M., Hussein, A.S., Tolba, M.F.: Spatio-Temporal motion detection for intelligent surveillance applications. *Int. J. Comput. Methods*, **11**(1) (2014)
5. Al-Berry, M.N., Salem, M.A.-M., Mousher, H.E., Hussein, A.S., Tolba, M.F.: Directional stationary wavelet-based representation for human action classification. In: *Advanced Machine Learning Technologies and Applications(AMLT2014)*, pp. 309–320. Cairo: Springer International Publishing (2014)
6. Bradley, A.P.: Shift-invariance in the discrete wavelet transform. In: *Proceeding of VIIIth Digital Image Computing: Techniques and Applications*, 1, pp. 29–38. Sydney (2003)

7. Bregonzio, M., Xiang, T., Gong, S.: Fusing appearance and distribution information of interest points for action recognition. *Pattern Recogn.* **45**, 1220–1234 (2012)
8. Daubechies, I.: Recent results in wavelet applications. *J. Electron. Imaging* **7**(4), 719–724 (1998)
9. Daubechies, I.: *Ten Lectures on Wavelets*. Industrial and Applied Mathematics Publishing Company, Philadelphia (1992)
10. Fadili, J., Starck J.L.: Numerical issues when using wavelets. In: Robert, M.A. (ed.) *Encyclopedia of Complexity and Systems Science*, vol. 3, pp. 6352–6368. Springer, New York (2009)
11. Fowler, J.E.: The redundant discrete wavelet transform and additive noise. *Signal Process. Lett.* **12**(9), 629–632 (2005)
12. Gonzalez, R.C., Woods, R.E.: *Digital Image Processing*, 3rd edn. Printice Hall, Upper Saddle River (2008)
13. Gupta, J.P., Singh, N., Dixit, P., Semwal, V.B., Dubey, S.R.: Human activity recognition using gait pattern. *Int. J. Comput. Vision Image Process.* **3**(3), 31–53 (2013)
14. Hu, M.-K.: Visual pattern recognition by moment invariants. *IEEE Trans. Inf. Theory* **8**(2), 179–187 (1962)
15. Kong, Y., Zhang, X., Hu, W., Jia, Y.: Adaptive learning codebook for action recognition. *Pattern Recogn. Lett.* **32**, 1178–1186 (2011)
16. Liu, S., Liu, J., Zhang, T., Lu, H.: Human action recognition in videos using hybrid features. In: *Advances in Multimedia Modeling, Lecture Notes in Computer Science*, vol. 5916, pp. 411–421 (2010)
17. Mallat, S., Hwang, W.: Singularity detection and processing with wavelets. *IEEE Trans. Inf. Theory* **38**(2), 617–643 (1992)
18. Nanni, L., Barhnam, S., Lumini, A.: Combining different local binary pattern variants to boost performance. *Expert Syst. Appl.* **38**, 6209–6216 (2011)
19. Pietikainen, M.: *Computer vision using local binary patterns*, Computational Imaging and Vision, vol. 40. Springer, Berlin (2011)
20. Poppe, R.: A survey on vision-based human action recognition. *Image Vis. Comput.* **28**, 976–990 (2010)
21. Rapantzikos, K., Tsapatsoulis, N., Avrithis, Y., Kollias, S.: Spatiotemporal saliency for video classification. *Sig. Process. Image Commun.* **24**, 557–571 (2009)
22. Salem, M.: *Multiresolution image segmentation*. Ph.D. Thesis, Department of Computer Science, Humboldt-Universitaet zu Berlin. Berlin, Germany (2008)
23. Schüldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: *International Conference on Pattern Recognition (ICPR2004)*. vol. 3, pp. 32–36. IEEE (2004)
24. Shao, L., Gao, R., Liu, Y., Zhang, H.: Transform based spatio-temporal descriptors for human action recognition. *Neurocomputing* **74**, 962–973 (2011)
25. Sharma, A., Kumar, D.K.: Moments and wavelets for classification of human gestures represented by spatio-temporal templates, pp. 215–226 (2004)
26. Sharma, A., Kumar, D.K., Kumar, S., McLachlan, N.: Wavelet directional histograms for classification of human gestures represented by spatio-temporal templates. In: *10th International Multimedia Modeling Conference MMM'04*, pp. 57–63. IEEE (2004)
27. Siddiqi, M.H., Ali, R., Rana, M.S., Hong, E.-K., Kim, E.s., Lee, S.: Video-based human activity recognition using multilevel wavelet decomposition and stepwise linear discriminant analysis. *Sensors* **14**(4), 6370–6392 (2014)
28. Song, K.-C., Yan, Y.-H., Chen, W.-H., Zhang, X.: Research and perspective on local binary pattern. *Acta Automatica Sinica* **39**(6) (2013)
29. Stark, J.-L., Fadili, J., Murtagh, F.: The undecimated wavelet decomposition and its reconstruction. *IEEE Trans. Image Process.* **16**(2) (2007)
30. Zhao, Y., Jia, W., Hu, R.-X., Min, H.: Completed robust local binary pattern for texture classification. *Neurocomputing* **106**, 68–76 (2013)

The 1st International Conference on Advanced
Intelligent System and Informatics (AISI2015),

November 28-30, 2015, Beni Suef, Egypt

M. Gaber, T.; Hassanien, A.E.; El-Bendary, N.; Dey, N.
(Eds.)

2016, XI, 536 p. 213 illus., 67 illus. in color., Softcover

ISBN: 978-3-319-26688-6