Weighted Directional 3D Stationary Wavelet-based Action Classification

Maryam N. Al-Berry[1], Mohammed A.-M. Salem [1], Hala M. Ebied [1], Ashraf S. Hussein[2], Mohamed F. Tolba[1]

[1]Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt.

[2]Faculty of Computer Studies, Arab Open University, Head Quarter, Kuwait.

maryam_nabil@cis.asu.edu.eg; salem@cis.asu.edu.eg; halam@cis.asu.edu.eg; ashrafh@acm.org; fahmytolba@gmail.com

Abstract

The aim of intelligent surveillance is to conceive reliable and efficient systems having the ability to detect moving objects in complicated real world scenes. These systems also, track the detected objects and analyze their actions and activities. Many applications are built on these operations such as advanced robotics and human computer interaction. This paper aims at proposing a framework for joint detection and recognition of human actions in a surveillance scenario. This objective is achieved in two main steps. First, a modern 3D stationary wavelet-based motion detection technique is used for detecting motion in the video sequence. The 3D technique fuses the spatial and temporal information achieving accurate detection results in real-world scene variations. The output of the detector is used to obtain a directional multi-scale representation for the action performed in the processed frames. This representation is described using local and global descriptors. The local descriptor combines the directional information contained in the wavelet coefficients in a weighted manner using an entropy value.The new local descriptor provides a discriminative local features for the human actions.The motion detection and the action recognition steps have been tested using benchmark datasets and compared to state-of-the-art methods.

Keywords: *Motion Detection, 3D Stationary Wavelet, Human Action Representation, Hue Invariant Moments, Local Binary Patterns, weighted directional wavelet LBP histogram, Entropy Value.*

1. Introduction

Nowadays, intelligent surveillance has become one of the most important applications, especially in security sensitive fields, such as banks and airports. Generally, surveillance may be defined as the observation of varying information, activities, or behaviors for some purposes. A typical framework of visual surveillance systems includes moving object detection, object classification, tracking, as well as behavior understanding and description [1]. The moving object detection module is responsible for segmenting moving objects from stationary background. This module is the base for any later processing; thus it must be accurate, robust and fast. Moving object detection techniques should be able to distinguish between variations in the pixel values that belong to motion and other variations caused by complexities in the scene, such as achange in lighting. After moving object detection, the detected objects are classified and the objects of interest are then processed for higher analysis, such as action and activity recognition.

There are many popular methods for detecting moving objects, with different performances, including frame difference [2], background subtraction [3], optical flow [4], and wavelet filters [5]. Frame differencing techniques are suitable for real-time applications, but when the moving object is very small or slowly moving it becomes hard to distinguish between real motion and noise spots. Wavelet filters vebeen proposed for the detection of very small low contrast objects from Infra Red images [5], but it required long computational time. Recently, multi-resolution techniques have been used in the motion detection process, including wavelets, which provides a way for analyzing signals with sharp changes in addition to the localized analysis of larger ones.

One of the most interesting computer vision fields that benefit from wavelets is the field of human action recognition. Wavelets can be 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 [6]. In global techniques, a holistic representation is formed using spatial and temporal information while in local techniques, local features are extracted. Global representations maintain global information about the described action but it requires high computational resources for processing video data. On the other hand, local representations focus the computations to the regions of interest only across the video sequence, and consequentlyuse limited computational resources. The advantages of both global and local representation techniques can be combined to obtain more robust representations for human actions [7]. The 3D Wavelet Transforms [8, 9, 10] 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 [11, 12, 13] or local [10] representations for the performed actions.

The main objective of this paper is to propose a framework for joint detection and recognition of human actions in a surveillance scenario. This objective is realized in two main stages. First, the 3D Stationary Wavelet Transform [8] is proposed for motion detection and the performance of the proposed technique is evaluated in surveillance scenarios. The proposed techniques provided a way to detect changes in the spatio-temporal domain and solved some of the challenges facing motion detection, such as, temporally varying illumination changes. The second stage was to use this proposed technique in the field of human action classification.

The 3D SWT technique is used for human action representation as it encodes the motion occurring in the processed volume of frames from different resolutions and generates motion energy images [11] that contain information for discriminating between different types of actions. Preliminary results of this proposal were promising. The directional multi-scale information contained in the wavelet co-efficient was then used to increase the accuracy of the classification process [12, 13]. More robust results were obtained by combining the proposed techniques with Local Binary Patterns [14]. The power of the 3D Stationary Wavelet Transform (SWT) [8] 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) [14], which is called the weighted directional wavelet LBP histogram. The extracted LBP histogram is weighted using the entropy value of the wavelet coefficients. The proposed local descriptor is combined with Hu invariant moments [15] that are extracted from the Multi-scale Motion Energy Images (*MMEI*) [13] to get benefit from the global information contained in the wavelet representation. The performance of the proposed method is verifiedby several experiments using Weizmann[16] and KTH [17]datasets.

In this manner, the 3D SWT can be used in boththe human action detection and the classification in an intelligent surveillance application. Figure 1 illustrates the proposed framework.

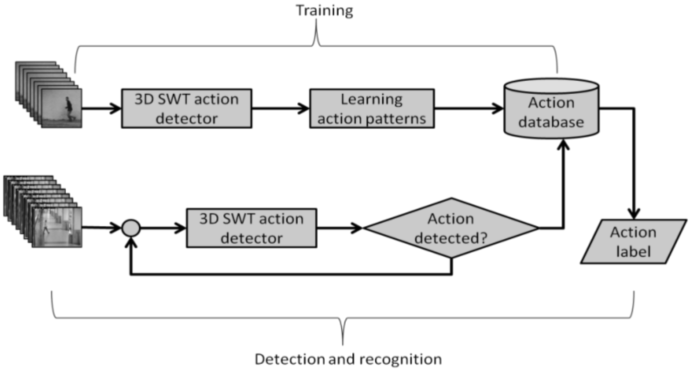

Figure 1. The proposed frame work for action detection and recognition

The rest of the paper is organized as follows: Section 2 reviews the background of the fields of motion detection and human action recognition. The proposed wavelet-based method is illustrated in section 3. In section 4, the 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 section 5.

2. Related Work

In this section the state–of –the–art is reviewed concerning the two main steps in the proposed framework, which are motion detection and human action recognition. This quick review focuses on wavelet-based techniques as they are the main interest of the paper.

2.1 Moving Object Detection

Motion detection is a special case of image segmentation with “changed” and “unchanged” regions. There are mainly three categories for video segmentation [9]: optical flow [4], temporal (frame) differencing[2], and background subtraction[18]. Wavelet filters were also used for motion detection, and to enhance other detection techniques.

Wavelet analysis [19, 20, 21] is a one of the most important mathematical disciplines that has gained a lot of interest in both theoretical and applied mathematics. The wavelet transforms are being used in different fields including image processing and computer vision as they provide the basis to describe signal features easily. The Decimated bi-orthogonal Wavelet Transform (DWT)[19] is probably the most widely used wavelet transform as it exhibits better properties than the Continuous Wavelet Transform (CWT). It has much fewer 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 [22].

One of the earliest attempts to use wavelet analysis in the motion detection was provided by Davies et al. [5], as they proposed to use one dimensional wavelet filters to detect very small objects, moving slowly, from a sequence of infra-red low contrast images. The idea was to replace differencing by taking the detail coefficient output from a Haar wavelet filter [20] applied temporally, at a fixed pixel location. The smallest known filter size combines two frames, and using larger filters enables combining more than two frames.

Two-dimensional DWT can be used to decompose an image into four sub-images, an approximation and three details. The approximation contains low frequency components, while the details contain high frequency components (fluctuations, edges, noise, etc.) in the horizontal, vertical, and diagonal directions. This capability to decompose the image into different frequency bands enables the motion detector to use useful frequencies and ignore irrelevant ones. In this manner, Cheng and Chen [23] proposed a method for detecting and tracking multiple moving objects, using the discrete wavelet transform. This method was successfully used in identifying the moving objects by their color and spatial information. The discrete wavelet transform was used to solve the problem of irrelevant motion in the background. Fang et al. [24] proposed a method for detection of moving cast shadows in traffic surveillance video. The values of the wavelet coefficients were used to differentiate between the pixels of cast shadow region and the pixels of the vehicle body.

Salem [9]introduced a 3D wavelet-based segmentation algorithm for extracting moving objects in a traffic monitoring application. The performance of this algorithm was compared to that of the 2D wavelet-based motion technique, and the comparison revealed that the 3D Wavelets have the advantage of considering movement in spatio-temporal domains, with the drawback of merging groups of moving objects.

Li et al. [25] proposed a robust pedestrian detection method in thermal infrared images. Their method is based on the Double-Density Dual-Tree Complex Wavelet transform (DD-DT CWT) and wavelet entropy. The Regions of Interest (ROIs) are located first, making use of high brightness property of the pedestrian pixels caused by the self-emission of the pedestrians related to the Planck’s law. The candidate ROIs were then decomposed by DD-DT CWT, and the wavelet entropy features are extracted from the high frequency sub-bands. The true pedestrian regions are finally classified and recognized, using the support vector machine (SVM) classifier. Crnojevic et al. [26] proposed a motion detection technique to detect temporal changes at multiple resolutions of stationary wavelet decompositions of input frames, and then fuse these partial detections to get the overall motion map. The authors used un-decimated wavelet transform [27], which is highly redundant but produces a stable set of wavelet coefficients suitable for temporal differencing. A 3D stationary wavelet-based motion detection technique has been proposed in [8]. The proposed technique fuses spatial and temporal analysis in a single 3D transform; and proved to be accurate and robust to real-world scene variations while having reasonable complexity.

2.2 Human Action Recognition

Automatic human action recognition is the process of analyzing the motion performed by a human and describing it using a name that can be understood by an average person[28, 6, 29]. An action is the ordered sequence of movements executed by a person to perform a specific task. The name given to this action is referred to as an action label. [30] The words “action” and “activity” are encountered in the literature, sometimes with the same meaning and sometimes to describe different levels of complexity in the recognition of movement. Different taxonomies and names of these levels of abstraction will be found in the literature. For example, Bobick [31] used “movement” for low level primitive actions, “activity” for a sequence of movements and “action” for high level events. Moeslund et al [32] and then

Poppe [6], used “action primitive” for atomicmovements, “action” for a set of action primitives that may describe whole body movements and “activities” for a number of subsequent actions. Turaga et al. [28] referred to simple motion patterns as “actions” and to complex sequences of actions executed by a number of possibly interacting humans as “activities”. Vishwakarma and Agrawal [33] categorized human activities into four levels of complexity. The first level is referred to as “gestures”: Motion of a part of the body in a very short time span such as; waving a hand. The second are “actions”: A single person activity composed of temporally ordered gestures such as running. “Interaction” is the third level referring to two or more persons performing an activity or a person interacting with an object, such as pointing a gun. Finally, “group activities” referred to activities performed by two groups of multiple objects such as groups of protesting people.

Wavelet analysis has been used in the context of 2D spatio-temporal action recognition. For example, Siddiqi et al. [34] 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). Their method focused on localized features from activity frames that discriminate their classes based on regression values. Sharma et al. [35] represented short actions using Motion History Images (MHI) [36]. They modified this representation to be invariant to translation and scale and used two dimensional, 3 level dyadic wavelet transforms but they concluded that the directional sub-bands alone were not efficient for action classification. In [37], Sharma and Kumar described the histograms of MHIs by orthogonal Legendre moments and modeled the wavelet sub-bands by Generalized Gaussian Density (GGD) parameters, shape factor and standard deviation. They found that the directional information encapsulated in the wavelet sub-band enhanced the classification accuracy.

Wavelet transforms have also been extended to 3D and 3D wavelet transforms [9, 10] have been proposed and used in computer vision applications. Rapantzikos et al. [10]used the 3D wavelet transform to globally represent dynamic events while keeping the computational complexity low. The wavelet coefficients were used to compute saliency and extract features. They treated the video sequence as a spatio-temporal volume and local saliency measures are generated for each visual unit (voxel) [10]. Shao et al. [38] applied a transform based technique to extract discriminative features from video sequences. They showed 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 [27], redundant wavelet [39] and shift invariant wavelet transform [40]. The SWT gives a better approximation than the DWT since it is linear, redundant and shift-invariant. In [8], the 3D SWT has been proposed and used for spatio-temporal motion detection. A 3D SWT is applied to each group of frames, and then the analysis is performed along the x-direction, the y-direction and the t-dimension of the time varying data [8], which is formed using the input frames.

Based on this 3D SWT, the authors of [11] proposed two spatio-temporal human action representations. The proposed representations were combined with Hue invariant moments as features for action classification. Results obtained using the public dataset were promising and provide a good step towards better enhancements. They proposed a directional global wavelet based representation of natural human actions [10] that utilizes the 3D SWT [8] 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) [14, 41] to describe texture images. LBPs are texture descriptors that have proved to be robust and computationally efficient in many applications [42, 38, 43, 44]. 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 [14]. The original LBP operator labels the pixels of an image by thresholding the 3x3 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 radiusR. This produces 2^Pdifferent binary patterns for the Ppixels 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 alack of directional information.

In this paper, the 3D SWT is used to detect the presence of motion in a sequence of frames. When motion is detected the wavelet coefficients are used to obtain a spatio-tempporal representation for the action performed. This representation is combined with LBPs to generate a new descriptor for human actions. 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 Wavelet-based Framework

Here, the two main steps that are used in the proposed framework are described in detail. First, the Motion detection step is described, and then comes the human action representation and description.

3.1 Motion Detection

The first main step in the proposed wavelet-based approach is motion detection. The proposed motion detection technique is based on the 3D Stationary Wavelet Transform. The idea is to benefit from the wavelet transform’s capabilities to highlight and localize variations in addition to fusing temporal analysis with the spatial analysis in one 3D transform. Multi-resolution analysis is applied and the detection results, from different scales, are fused to improve the motion detection process. The proposed technique is comprised of two main steps: motion detection and binarization. The motion detection step includes preparing the block of the time varying data to be processed, applying the 3D SWT, and fitting the selected detail images to a normal distribution with zero means and small standard deviation. The binarization step is carried out by thresholding the outliers, ORing the obtained motion masks, applying a morphological opening, and finally spatial median filtering.

First, the video sequences are divided into frame groups; the number of frames in each group depends on the required number of levels in the analysis. A 3D SWT is applied to each group of frames. The analysis is then carried out along the x-direction, the y-direction, and the t-direction of the time varying data, which is formed using the input frames. Single level 3D transform, shown in Figure2, is implemented as three 1D stationary wavelet transformations.

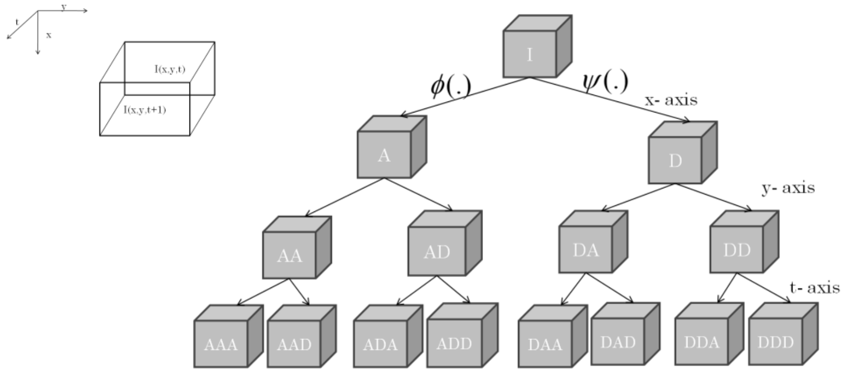

Figure 2. Illustration of one level 3D Stationary wavelet Transform

The proposed algorithm starts with computing the 3D Stationary Wavelet Transform (SWT) coefficients $w_j^d(x,y,t)$, where j is the resolution (3 resolutions were used in the analysis), and d is the sub-band orientation (Horizontal=1, Vertical=2, Diagonal=3) and the approximation coefficient $c_j(x,y,t)$ computed by the associated scaling function.

To compute these coefficients, the "à torus" algorithm [45]can be extended to 3D, by the same way used for the 2D in [9], as follows:

$$c_{j+1}[k,l,m] = (\bar{h}^{(j)}\bar{h}^{(j)}\bar{h}^{(j)} * c_j)[k,l,m] \tag{1}$$
$$w_{j+1}^1[k,l,m] = (\bar{g}^{(j)}\bar{h}^{(j)}\bar{h}^{(j)} * c_j)[k,l,m] \tag{2}$$
$$w_{j+1}^2[k,l,m] = (\bar{h}^{(j)}\bar{g}^{(j)}\bar{h}^{(j)} * c_j)[k,l,m] \tag{3}$$
$$w_{j+1}^3[k,l,m] = (\bar{g}^{(j)}\bar{g}^{(j)}\bar{h}^{(j)} * c_j)[k,l,m] \tag{4}$$
$$w_{j+1}^4[k,l,m] = (\bar{h}^{(j)}\bar{h}^{(j)}\bar{g}^{(j)} * c_j)[k,l,m] \tag{5}$$
$$w_{j+1}^5[k,l,m] = (\bar{g}^{(j)}\bar{h}^{(j)}\bar{g}^{(j)} * c_j)[k,l,m] \tag{6}$$
$$w_{j+1}^6[k,l,m] = (\bar{h}^{(j)}\bar{g}^{(j)}\bar{g}^{(j)} * c_j)[k,l,m] \tag{7}$$
$$w_{j+1}^7[k,l,m] = (\bar{g}^{(j)}\bar{g}^{(j)}\bar{g}^{(j)} * c_j)[k,l,m] \tag{8}$$

where g and h are the analysis filters of the wavelet function, and the associated scaling function, respectively, $h[n]$, $g[n]$ are the impulse responses of the filters h, g, $\bar{h}[n] = h[-n],\quad \bar{g}[n] = g[-n],\quad n \in \mathrm{Z}$ are their time reversed versions.

The motion is highlighted in the temporal changes that happened along the t-axis, which are represented in the detail sub-images ADD, DAD, and DDD, and denoted $w_j^d(x,y,t)$, d = (5,6 and 7). The obtained wavelet coefficients at each resolution and orientation are assumed to be normally distributed with zero means and small standard deviation. The moving objects in the scene will cause some wavelet coefficients to change and become outliers to the normal distribution, and so the problem of finding moving pixels is considered as detecting outliers in the set of wavelet coefficients. Following [26]a modified z-score $z_j^d(x,y,t)$ is computed and compared to a threshold τ. The modified z-score is calculated by normalizing the wavelet detail coefficients with the median of absolute deviation (MAD) about the median, as follows:

$$MAD(w_j^d(x,y,t)) = median_{x,y}\left\{\left|w_j^d(x,y,t)\right|\right\} \qquad (9)$$

Modified z-score is computed as:

$$z_j^d(x,y,t) = 0.6755\frac{w_j^d(x,y,t)}{MAD(w_j^d(x,y,t))} \qquad (10)$$

Assuming that moving objects occupy a small part of the scene, modified z-score test results in a motion detection mask that identifies moving pixels.

$$O(x,y,t) = \begin{cases} 1 & if\ \left|z_j^d(x,y,t)\right| > \tau \quad for\ some \quad j,d \\ 0 & \quad otherwise \end{cases} \qquad (11)$$

The threshold value τ can be determined manually using Receiver Operator Characteristic (ROC) curves or using an automatic technique that depends on the image statistics, such as Median Absolute Deviation (MAD) thresholding described in [46]. A foreground image is obtained by logical ORing with the different motion detection masks, obtained from the different sub-bands and resolutions. This means that a pixel is identified as foreground pixel if it is identified as a moving pixel in at least one of the motion detection masks. After obtaining the foreground image, a morphological open operation is applied followed by 3×3 median filter to obtain the final foreground image. Morphological operations are used to improve the foreground mask by eliminating small false-positive or false-negative regions, and merging nearby disconnected foreground regions.

3.2 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 Figure 3.

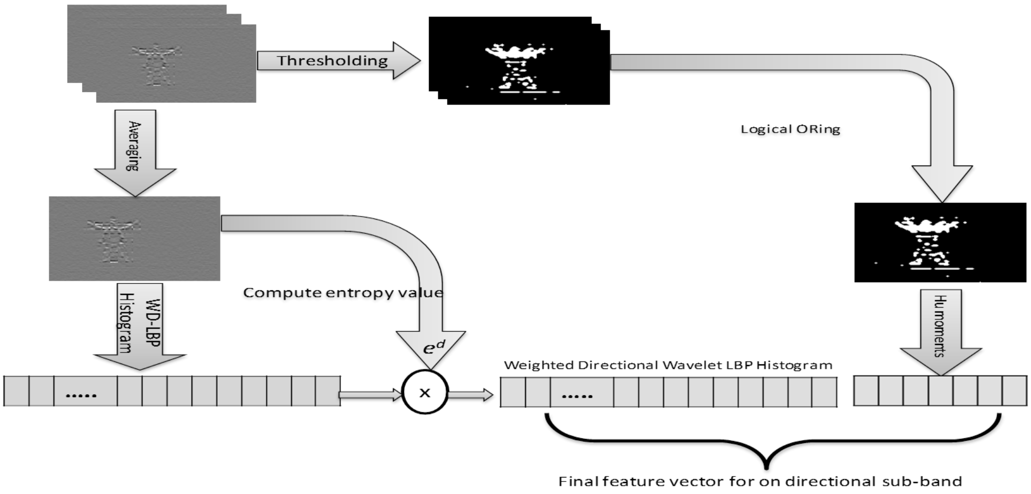

Figure 3 Illustration of the proposed method

To obtain on the proposed descriptor, the video sequence is treated as a 3D volume by considering time the third dimension and is processed in several steps as follows:

First, the video sequence is divided into blocks of frames and these blocks are supplied to the 3D SWT to obtain 3 levels of spatio-temporal coefficients $w_j^d(x,y,t)$, where *(x,y)* are the spatial co-ordinates of the frames, t is the time, j is the resolution and d is the sub-band orientation, and the approximation coefficient $c_j(x,y,t)$ computed by the associated scaling functions. The detail sub-bands $w_j^d(x,y,t), d = (5,6,7)$ are for representing the action as they contain highlighted motion detected in the temporal changes that happened along the t-axis. The resulting coefficients are used to obtain a directional multi-scale stationary wavelet representation for the action by averaging the wavelet coefficients of the three obtained resolutions. This results in three Directional multi-scale Stationary Wavelet Templates $\mathrm{DSWT^d}$, $\mathrm{d} = (5,6,7)$, for the three detail sub-bands.

$$\mathrm{DSWT^d(x,y)} = \frac{\sum_{\mathrm{j=1}}^{3} \mathrm{w_j^d\,(x,y)}}{3}, \mathrm{d} = (5,6,7) \tag{12}$$

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

$$DW - LBP_{(P,R)}^{d} = \sum_{p=0}^{P-1} S(f_p^d - f_c^d)2^p \tag{13}$$

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

$$S(z) = \begin{cases} 1, & z \geq 0 \\ 0, & z < 0 \end{cases} \tag{14}$$

f_c^d is the center pixel in the defined neighborhood in sub-band d, f_p^d , $p = 0,1,\cdots,P-1$ 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 [47] for each directional wavelet template (e^d) is computed and used to give a weight for the computed histogram.

The coefficients of the three selected sub-bands are thres holded to obtain the motion images. The motion images obtained at the eighth layer of the 3 different scales are fused into three directional multi-scale motion energy images (*MMEI^d(x,y,t)*). This multi-scale motion energy image for each directional sub-band *d* (*MMEI^d*) is obtained by logical ORing the motion images obtained from different resolutions for this sub-band. These sub-band motion energy images encode the directional motion energy during the processed 8 frames at 3 different scales and thus can be used to represent the action in the duration of these 8 frames. These coefficients of the different scales are threshold to obtain motion images $O_j^d(x,y,t)$

$$O_j^d(x,y,t) = \begin{cases} 1 & if\ w_j^d(x,y,t) > \tau\ for\ some\ j,d \\ 0 & otherwise \end{cases} \tag{15}$$

The threshold value τ can be determined automatically or empirically. In this chapter, Otsu's thres holding technique [47] is used to determine the suitable threshold value automatically. This technique maximizes the inter-class variance and is entirely based on the computation of the image histogram [47]. To describe the obtained energy image the seven Hu moments [15] are computed. The Hu moments are known to give a good global description for shapes while being a translation, scale, mirroring and rotation invariant [47].

4. Results and Discussions

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

4.1 Human Action Datasets

The performance evaluation of the proposed method is carried out using the Weizmann [16] and KTH dataset[17]. Weizmann dataset [16] contains samples for 10 actions, performed by nine persons. The background of the scenes is static. Fig. 5 shows some sample frames from the dataset.

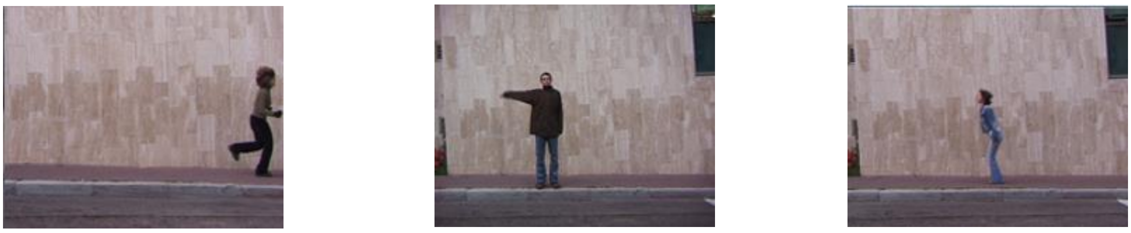

Figure 4. Sample frames from Weizmann dataset.

KTH dataset contains four different scenarios s1, s2, s3, and s4 [17]. Each scenario contains samples for six different actions performed by 25 different people. The first scenario s1 is “outdoors”, s2 is “outdoors with scale variation”, s3 is “outdoors with different clothes”, and s4 is “indoors”. This dataset is widely used by researchers, as it is considered one of the largest single view datasets with respect to the number of sequences. Different sample frames and scenarios are shown in Figure5.

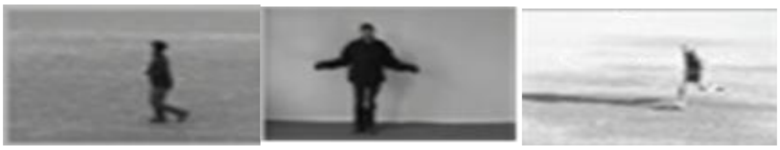

Figure 5 Sample of KTH dataset actions and scenarios.

4.2 Experiments and Results

A simple K-Nearest Neighbor (KNN) classifier and Decision Tree (DT) classifier were used for testing the accuracy of the proposed features. The KNN classifier uses Euclidean 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 in 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.

Two classification schemes were used in the evaluation. First, the features of the three directional sub-bands are combined into 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 Weizmann dataset.

Table1. Performance obtained on Weizmann dataset

Band	KNN%	DT%
w^5	90.32	89.25
w^6	86.02	86.02
w^7	87.10	90.32
Average	87.81	88.53
Sub-band Feature Fusion	86.02	89.25
Voting	90.32	89.25

The four different scenarios of the KTH database were examined separately; the overall performance is set to be the average of the performances of the four scenariosResults obtained on KTH dataset are shown in table 2. 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.

Table 2. Accuracy using combined feature vector vs. voting between directional feature vectors on KTH dataset

Scenario	**Combined feature vector %**		**Voting %**	
	KNN	DT	KNN	DT
s1	94.67	92.67	95.33	98.00
s2	94.67	94.67	92.00	94.67
s3	94.67	94.67	94.67	96.00
s4	96.00	96.67	96.00	96.67
Average	95.00	94.67	94.50	96.34

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 3.

Table 3. Accuracy using combined feature vector vs. voting for gait actions in KTH dataset

Scenario	**Combined Feature Vector %**		**Voting %**	
	KNN	DT	KNN	DT
s1	97.33	94.67	100	100
s2	96.00	96.00	100	100
s3	98.67	98.67	100	100
s4	97.33	97.33	100	100
Average	97.33	96.67	100	100

Table 4 shows a comparison between the proposed method and existing state-of-the-art methods. The proposed method outperformed existing techniques, achieving higher classification accuracy.

Table Error! No text of specified style in document.**4. Performance Comparison of Existing Techniques**

Method	Accuracy %
Proposed method (CFV)	95.00
Proposed method (voting)	94.50
Kong et al. (2011) [48]	88.81
Bregonzio (2012) [49]	94.33
Gupta et al. (2013) [50] (gait actions Only)	95.10
Proposed method (gait actions only)(CFV)	97.33
Proposed method (gait actions only)(voting)	100

6. Conclusions and Future Work

The area of intelligent surveillance has gained a lot of interest in the computer vision community. The ultimate goal of the intelligent surveillance is to build a robust system that is capable of detecting, tracking, and identifying objects in complicated real world conditions. The system should also be able to recognize different actions performed by the objects.

This paper gives a brief review of recent state-of-the art in the area of using wavelets and mulit-resolution analysis in the different stages of intelligent surveillance. The paper also proposes a wavelet-based framework for using 3D multi-scale stationary wavelet analysis for human action recognition in a surveillance scenario. The proposed framework uses the 3D multi-scale stationary wavelet transform to detect the presence of moving objects in the scene. For the detected moving objects the output of the 3D stationary wavelet transform is used as a spatio-temporal representation of the action performed during the processed frames. Local binary patterns are then used to describe local structures in the wavelet coefficients. 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 standard classifiers. 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 investigating the effect of using different wavelet functions and investigating the effect of increasing the number of analysis levels in the 3D transform. The proposed description can be also extended the 3D.

References

[1]. W. Hu, T. Tan, L. Wang and S. Maybank, "A survey on visual surveillance of Object Motion and Behaviors". IEEE Transaction on Systems, Man, and Cybernitics-Part C: Applications and Reviews, 34(3), 2004, pp. 334-352.

[2]. Collins, Lipton, Kanade, Fujiyoshi, Duggins, Tsin, Tolliver and Hasegawa, "A System for video surveilance and monitoring". Final Report of the VSAM project. Technical report CMU-RI-TR-00-12, Robotics Institute, Carnegie Mellon University, May..

[3]. C. Poppe, G. Martens, P. Lambert and R. V. Walle, "Mixture models based background subtraction for video surveillance applications". In Proceedings of the 12th International Conference on Computer Analysis of Images and Pattern. August

27–29; Vienna, Austria. 2007, pp. 28–35.

[4]. U. Knauer, T. Dammeier and B. Mefffert, "The structure of road traffic scenes as revealed by unsupervised analysis of the time avaraged optical flow". In Proceedings of 17th International Conference on the Applications of Computer Science and Mathematics in Architecture and Civil Engineering. July 12–14. Weimar, Germany, 2006.

[5]. D. Davies, P. Palmer and M. Mirmehdi, "Detection and Tracking of Very Small Low Contrast Objects".In Proceedings of British Machine Vision Conference, Southampton, UK, 1998.

[6]. R. Poppe, "A survey on vision-based human action recognition," Image and Vision Computing, vol. 28, 2010, pp. 976-990.

[7]. S. Liu, J. Liu, T. Zhang and H. Lu, "Human action recognition in videos using hybrid features".In Advances in Multimedia Modeling, Lecture Notes in Computer Science, vol. 5916, 2010, pp. 411-421.

[8]. M. N. Al-Berry, M. A.-M. Salem, A. S. Hussein and M. F. Tolba, "Spatio-temporal motion detection for intelligent surveillance applications". International Journal of Computational Methods, 12(1), 2015.

[9]. M. Salem, "Multiresolution Image Segmentation".Humboldt University, Berlin, Germany, 2008.

[10]. K. Rapantzikos, N. Tsapatsoulis, Y. Avrithis and S. Kollias, "Spatiotemporal saliency for video classification". Signal Processing: Image Communication, vol. 24, 2009, pp. 557-571.

[11]. M. N. Al-berry, M. A.-M. Salem, H. M. Ebeid, A. S. Hussein and M. F. Tolba, "Action recognition using stationary wavelet-based motion images".In IEEE conference on Intelligent systems 14, Warasaw, Poland, 2014, pp. 743–753.

[12]. M. N. Al-Berry, M. A.-M. Salem, H. E. Mousher, A. S. Hussein and M. F. Tolba, "Directional stationary wavelet-based representation for humanaction classification".In Advanced Machine Learning Technologies and Applications(AMLTA2014), Cairo, 2014, pp. 309–320.

[13]. M. N. Al-Berry, H. M. Ebied, A. S. Hussein and M. F. Tolba, "Human action recognition via multi-scale 3D stationary wavelet analysis".In Hybrid Intelligent Systems 2014 (HIS'14), Kuwait, 2014.

[14]. Pietikainen, "Computer Vision Using Local Binary Patterns". Computational Imaging and Vision, vol. 40, Springer-Verlag, 2011.

[15]. M.-K. Hu, "Visual pattern recognition by moment invariants". IEEE Transacions on Information Theory, 8(2), 1962, pp. 179-187.

[16]. M. Blank, L. Gorelick, E. Shechtman, M. Irani and R. Basri, "Actions as space-time shapes".In International conference on Computer Vision,Beijing, China, 2005, pp. 1395–1402.

[17]. C. Schüldt, I. Laptev and B. Caputo, "Recognizing human actions: alocal SVM approach".In Int. Conf. Pattern Recognition, Washington, DC, USA, 2004, pp. 32–36.

[18]. C. R. del-Blanco, F. Jaureguizar and N. García, "An efficient multiple object detection and tracking framework for automatic counting and video surveillance applications". IEEE Transactions on Consumer Electronics, 53(3), 2012, pp. 857-862.

[19]. I. Daubechies, "Recent results in wavelet applications". Journal of Electronic Imaging, 7(4), 1998, pp. 719-724.

[20]. I. Daubechies, "Recent results in wavelet applications". Journal of Electronic Imaging, 7(4), 1998, pp. 719-724.

[21]. S. Mallat, W. L.Hwang, "Singularity detection and processing with wavelets". IEEE Trans. Inf. Theory , 38(3), 1992, pp. 617–643

[22]. J. -L.Stark, J. M. Fadili, "Numerical issues when using wavelets". In: Meyers and Robert (Eds.), Encyclopedia of Complexity and Systems Science. Springer, New York, 2009, pp. 6352–6368.

[23]. F.-H. Cheng and Y.-L. Chen, "Real-time multiple object tracking and identification based on discrete wavelet transform". Pattern Recognition, 39(6), 2006, pp. 1126-1139.

[24]. L. Z. Fang, W. Y. Qiong and Y. Z. Sheng, "A method to segment moving vehicle cast shadow based on wavelet transform". Pattern Recognition Letters, vol. 29, 2008, pp. 2182–2188.

[25]. J. Li, W. Gong, W. Li and X. Liu, "Robust pedestrian detection in thermal infrared imagery using the wavelet transform". Infrared Physics & Technology, vol. 53, 2010, pp. 267-273.

[26]. V. Crnojevic, B. Antic and D. Culibrk, "Optimal wavelet differencing method for robust motion detection".In Proceedings of International Conference on Computer Engineering and Systems ICCES'2009, Cairo, Egypt, 2009.

[27]. J.-L. Stark, J. Fadili and F. Murtagh, "The undecimated wavelet decomposition and its reconstruction". IEEE Transactions on Image Processing, 16(2), 2007, pp. 297–309.

[28]. P. Turaga, R. Chellappa, V. Subrahamanian and O. Udrea, "Machine recognition of human activities: a survey". IEEE Transactions on circuits and systems for video technology, 18(11), 2008, pp. 1473-1487.

[29]. F. Lv, R. Nevatia and M. Wai Lee, "3D Human action recognition using spatio-temporal motion templates". Computer Vision in Human-Computer Interaction Lecture Notes in Computer Science, vol. 3766, 2005, pp. 120 - 130.

[30]. D. Weinland, R. Ranford and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition". Computer Vision and Image Understanding, vol. 115, 2011, pp. 224-241.

[31]. A. F. Bobick, "Movement, activity, and action: The role of knowledge in the perception of motion". Philosoph Trans Roy Soc Lond B, vol. 352, 1997, pp. 1257-1265.

[32]. T. B. Moeslund, A. Hilton and V. Kruger, "A survey of advances in vision-based human motion capture and analysis". Computer Vision and Image Understanding, vol. 104, 2006, pp. 90–126.

[33]. S. Vishwakarma and A. Agrawal, "A survey on activity recognition and behavior understanding in video surveillance". The Visual Computer, 29(10), 2013, pp. 983-1009.

[34]. M. H. Siddiqi, R. Ali, M. S. Rana, H. E-K, K. E. S and L. S, "Video-based human activity recognition using multilevel wavelet decomposition and stepwise linear discriminant analysis". Sensors, 14(4), 2014, pp. 6370-6392.

[35]. A. Sharma, D. K. Kumar, S. Kumar and N. McLachlan, "Wavelet directional histograms for classification of human gestures represented by spatio-temporal templates".In 10th International Multimedia Modeling Conference MMM'04, 2004, pp. 57–63.

[36]. M. Ahad, J. Tan, H. Kim and S. Ishikawa, "Motion history image: its variants and applications". Machine Vision and Applications, vol. 23, 2012, pp. 255-281.

[37]. A. Sharma and D. K. Kumar, "Moments and wavelets for classification of human gestures represented by spatio-temporal templates".In Advances in Artificial Intelligence, Springer Berlin Heidelberg, 2004, pp. 215–226.

[38]. L. Shao, R. Gao, Y. Liu and H. Zhang, "Transform based spatio-temporal descriptors for human action recognition". Neurocomputing, vol. 74, 2011, pp. 962-973.

[39]. J. E. Fowler, "The redundant discrete wavelet transform and additive noise". Signal Processing Letters, 12(9), 2005, pp. 629-632.

[40]. A. P. Bradley, "Shift-invariance in the Discrete Wavelet Transform".In Proceedings of VIIth Digital Image Computing: Techniques and Applications, Sydney, 2003.

[41]. K.-C. Song, Y.-H. Yan, W.-H. Chen and X. Zhang, "Research and perspective on localbinary pattern". Acta Automatica Sinica, 2013, 39(6).

[42]. Y. Zhao, W. Jia, R.-X. Hu and H. Min, "Completed robust local binary pattern for texture classification". Neurocomputing, vol. 106, 2013, pp. 68-76.

[43]. S. H. Salah, H. Du and N. Al-Jawad, "Fusing local binary patterns with wavelet features for ethnicity identification," International Journal of Computer, Information, Systems and Control Engineering, 7(7), 2013, pp. 347-353.

[44]. Y. Z. Goh, A. b. Teoh and M. K. Goh, "Wavelet local binary patterns fusion as illuminated facial image preprocessing for face verification". Expert Systems with Applications, vol. 38, 2011, pp. 3959–3972.

[45]. M. J. Shensa, "The Discrete Wavelet Transform : Wedding the A trous and Mallat Algorithms". IEEE Transactions on Signal Processing, 40(10), 1992, pp. 2464-2482.

[46]. P. L. Rosin and T. Ellis, "Image difference threshold strategies and shadow detection".In Proceedings of the 6th British Machine Vision Conference, Birmingham, UK, 1995.

[47]. R. C. Gonzalez and R. E. Woods, "Digital Image Processing". Third ed., Printice Hall, 2008.

[48]. Y. Kong, X. Zhang, W. Hu and Y. Jia, "Adaptive learning codebook for action recognition". Pattern Recognition Letters, vol. 32, 2011, pp. 1178–1186.

[49]. M. Bregonzio, T. Xiang and S. Gong, "Fusing appearance and distribution information of interest points for action recognition".Pattern Recognition, vol. 45, 2012, pp. 1220–1234.

[50]. J. P. Gupta, N. Singh, P. Dixit, V. B. Semwal and S. R. Dubey, "Human Activity Recognition using Gait Pattern," International Journal of Computer Vision and Image Processing, 3(3), 2013, pp. 31–53.