Application of 3D-Wavelet Statistics to Video Analysis

M. Omidyeganeh[1,2,4], S. Ghaemmaghami[1,3], S. Shirmohammadi[4]

[1]Electrical Engineering Department, [2]Advanced Information & Communication Technology Center (AICTC),

[3]Electronics Research Institute, Sharif University of Technology

[4]Distributed and Collaborative Virtual Environment Research Lab., University of Ottawa

m_omid@ee.sharif.edu, ghaemmag@sharif.edu, shervin@discover.uottawa.ca

ABSTRACT

Video activity analysis is used in various video applications such as human action recognition, video retrieval, video archiving. In this paper, we propose to apply 3D wavelet transform statistics to natural video signals and employ the resulting statistical attributes for video modeling and analysis. From the 3D wavelet transform, we investigate the marginal and joint statistics as well as the Mutual Information (MI) estimates. We show that marginal histograms are approximated quite well by Generalized Gaussian Density (GGD) functions; and the MI between coefficients decreases when the activity level increases in videos. Joint statistics attributes are applied to scene activity grouping, leading to 87.3% accurate grouping of videos. Also, marginal and joint statistics features extracted from the video are used for human action classification employing Support Vector Machine (SVM) classifiers and 93.4% of the human activities are properly classified.

Keywords — Video analysis, 3D wavelet transform statistics, Human action recognition.

1. Introduction

Video today has an important role in the transmission of information to a wide range of users in various applications. With the ever increasing availability of both processing power and bandwidth, whether in desktop or mobile settings, video applications and services, such as those offered by YouTube, Google Video, and many others, are becoming more ubiquitous and part of everyday life. At the same time, users' expectations are increasing too and more intelligent and interactive features are needed. One of those features is human action recognition which has become a popular field of research [1-18] in video modeling and analysis, and can be used in human activity analysis, gesture recognition, biometrics [2], video indexing and retrieval, surveillance systems. In human action recognition, features of the signal are extracted from video sequences and are used to determine the actions. Accordingly, the extracted features and the classification method significantly affect the performance and efficiency of these systems.

At a high level, video modeling and analysis can contribute to and improve the above techniques. Video analysis extracts the video signal parameters that convey critical characteristics of the signal. A critical characteristic can be, for example, a scene change, a salient frame, a moving object or a specific event. Video modeling and analysis provide appropriate means to process the signal and mine necessary information in order to get the desired output. For example, in video compression, video modeling helps to detect the main parts of the signal, such as the foreground or the moving objects, to assign more bits (higher quality) to those parts of the information. Video modeling and analysis can therefore enhance the coding efficiency of the video and lead to a better compressed signal. In addition to encoding, video retrieval can also benefit from video modeling; for instance, the retrieval rate can be improved by extracting proper parameters from the signal and constructing a suitable feature vector along with an appropriate distance measure. Finally, the model of the video signal and its important parameters can be used in source separation applications, where the original signal has specific characteristics and the known set of characteristics can be utilized to separate it from a mixed signal such as the combination of the original signal and the noise.

In this paper, our main contribution is a new method for modeling and analysis of natural videos based on the statistical properties of the 3D-wavelet transformed video signal. To the best of our knowledge, this is the first approach that marginal and joint statistics of 3D wavelet transform are investigated and used as features that are shown to be better indications of the human interpretation of video contents, as compared to the existing methods for video modeling and analysis. We demonstrate this efficiency by deploying our method in two applications: scene activity grouping and human action recognition. In the latter, our method leads to a high accuracy of 93.4% in the classification of the KTH human action database [18], outperforming existing methods. Furthermore, we suggest a new definition for activity level in a given video. The activities are categorized into slow or fast motion, depending on the speed of the changes in the time domain, and are identified as local or global, based on the relative fraction of the frame involved in the activity. Thus, four activity levels are proposed. We then use the features extracted from the joint distribution functions, to attain information about activity level in videos. This information is used to group videos into four different activity sets with 87.3% accuracy.

The rest of this paper is structured as follows. We continue by taking a look at the related work in this field in section 2. A brief explanation of the 3D wavelet and relationships between its coefficients is given in section 3. In section 4, the marginal and joint statistics of the wavelet transform of natural video signals are extracted and studied , while in section 5 the MI is estimated as a quantitative feature and is used as a measure of dependency between coefficients and their parents, cousins, or neighbors. This section also describes the association between these estimates and the video activity level. Section 6 contains the experimental results, where we use 3D wavelet marginal and joint statistical features in human action recognition. The relationship between the kurtosis graphs and the type and the amount of the video activity is also discussed in this section. Finally, the paper is concluded in section 7.

2. Related Work

2.1. Video Content Analysis

Video modeling and video analysis have been of great interest in the video research community. In the existing literature, video analysis parameters are typically used to analyze video contents [19, 20]. In [21], a one-dimensional representation of frames has been introduced using Mojette transform which is the discrete form of the Radon transform. This idea is used in motion estimation, scene change detection, and region of interest extraction. In another work, shots are recognized by one-dimensional Mosaics based on X-ray projections of each frame corresponding to the total of pixel values in both vertical and horizontal directions [22]. Applying this transform to the video sequence, the portion of the frame related to the background is segmented based on the motion estimation. The main problem with these methods, however, is the difficulty in the selection of a proper local similarity measure and the window size to calculate this measure to track both short and long term changes, since in most video processing applications, the ability to track the temporal information helps to improve the system efficiency.

Another approach to the expression of the temporal interrelations of video contents is to utilize the temporal features, such as optical flow [19] – as a dense field of motion representation- and motion vectors [23, 24]. A local descriptor as well as the optical flow is used to exhibit regions of interest and temporal information in [19]. In [23], motion fields are taken as separate signals resembling time series. In another work [20], spatio-temporal slices are employed to present motion patterns and extract key frames. These methods, however, can hardly reveal and quantify long term temporal relations between video contents.

For applications such as shot classification, video retrieval, and video indexing, accurate information about temporal evolution of the video signal is quite useful. Time series modeling algorithms can be used to model temporal associations of the sequences of the spatial features. Statistical analysis is one of the basic approaches to temporal modeling [25-30]. In [25], a layered Hidden Markov Model (HMM) is introduced to do video semantic analysis. A Markov Chain Monte Carlo (MCMC) based algorithm is used in [26] to model video scene segmentation. Scene boundaries are selected based on MCMC. The initial locations of boundaries are selected randomly and updated automatically in the procedure of MCMC. In [27] a hidden Markov tree is employed to model relationships between coefficients in 2D wavelet transform. Markov fields are also used in scene modeling [28]. The model combines the visual information and camera motion data and employs the resulting vector to investigate the changes over time and studies lossy and lossless information rates based on the achieved dynamic model and find conditions for tight bounds. Markov models efficiently characterize complicated evolutional systems and relations, though appropriate assumptions about statistical distributions and accurate computations are needed to reach their target fitting models [29]. In [30], Auto-Regressive (AR) models, which are simplified linear Markov models, are used to model video evolution based on color histograms features. This method models the temporal evolution of successive video frames and extracts keyframes and shots; however, employing more appropriate features corresponding to the Human Visual System (HVS) characteristics improves the results [56].

In another analysis scheme, the video signal is taken as a three-dimensional signal to reach a fitting statistical model. A Gaussian Mixture Model (GMM) maps the video pixels from a 3D space-time domain to a 7D feature space and segments the video into main objects [31, 32]. The results are used in shot extraction, key frame selection, video editing, motion detection, and event detection. To avoid time delay and reduce computational problems, a piecewise GMM is presented in [32], while there are still considerable temporal and spatial redundancies left unused. 3D wavelet transform has been applied to detect scene changes in the video sequence in [33]. First, 2D transform of each frame is calculated. Next, 1D transform is applied to the 8-frame length temporal evolution of each coefficient in time, and then the correlation between adjacent frames and three other simple features, extracted from the 3D wavelet transform, is used to detect scene changes.

In this work we study the statistical properties of 3D wavelet transform of video signals. Marginal and joint statistics and MI estimates between wavelet transform coefficients are investigated and utilized to analyze the video signal. We have employed the statistical features to scene activity classification and human action recognition. The high efficiency achieved by the proposed method can be attributed to utilizing two important factors: 1) employing the wavelet transform which has the ability to attain spatio-temporal attributes of the video signal, yielding a sparse representation of the signal, and matching with the frequency distribution function of the Human Visual System (HVS) [34, 35]; and 2) Choosing appropriate marginal and joint wavelet features which convey motion information based on HVS [27] characteristics.

2.2. Human Action Recognition

Human action recognition is an active field of research in video analysis and modeling and its results can be utilized in many applications, such as human activity analysis, gesture recognition, biometrics [2], video indexing and retrieval, surveillance systems. In [4] a hierarchical concept is proposed to human movement classification: an action primitive – each small movement of the limb-, an action – a succession of action primitives that cause the whole body movement- and an activity – a combination of actions. In this paper, our definition of “action” is the same as described in [1, 3, 4] which is a combination of simple patterns of motion performed by one person in a video sequence; e.g., ‘walking’ or ‘boxing’. The main problem in human action recognition systems is suitable feature extraction from video sequences, which then reduces the problem to classification. There are two main approaches to feature selection [3]: global and local image representations.

The global approach considers the whole human body as the region of interest and extracts features from this region. It employs rich information, achieves excellent results, and the feature extraction procedure is uncomplicated. But this method requires exact background extraction or body tracking techniques and is sensitive to noise and occlusion [5, 6]. To resolve this concern a grid-based scheme, another version of this method, can be employed [7], in which the desired region is divided into cells spatially and each cell is encoded locally. A group of frames can be used together to form three dimensional cells and have spatio-temporal descriptors [8]. However there is still a need to have a global overview of the body in these schemes.

The local approach is to utilize local descriptors as a combination of independent patches calculated around interest points [9-11], which are the points in which abrupt changes occur in the spatial or temporal domains since they have more information than other points in the sequence. Patches then construct the bag-of-features. This approach is more robust against foreground variations, noise, and partial occlusion but needs a pre-processing step. As a result, the comparison of two video sequences will not be simple and the patches are always clustered and codebooks calculated from the patch clustering are employed to represent a video sequence as a bag of features. However there is always redundancy in the extracted data. Some works consider spatial and temporal correlations between patches to reduce the redundancy [12, 14, 15].

We employ 3D wavelet transform for human action recognition. The proposed method can be considered as a global image representation approach to human action recognition, since it extracts global descriptors from video signals. We apply the wavelet transform to the difference of adjacent frames, as this transform is sensitive to the edges and their variations in time. Therefore, our method does not need any background elimination step and recognizes the actions regarding the variations of edges in time. Our contribution is to investigate the 3D wavelet statistical properties, including marginal and joint statistical attributes as quality study, and MI as quantity study, and to build on this investigation to propose a method to infer the activity level of the video to improve the efficiency of video processing applications such as human action recognition.

3. 3D Wavelet Transform

The Fourier transform decomposes a signal into its frequency components, whereas the cosine transform presents superior estimate of the signal with fewer coefficients. The Fourier transform is suitable for stationary signals and is not optimal in non-stationary cases; it gives global information about a signal which is not sufficient in many signal processing applications [37]. The wavelet transform represents a signal as a superposition of a set of basic functions [38]. This transform, unlike the cosine transform, can be applied to larger block sizes of the signal and thus overcomes the blocking problem. It provides sparse representation of signals, especially one-dimensional signals. Also, it gives a higher compression ratio, as compared to the cosine transform, and structurally conforms to the human perceptual system [34, 35]. The wavelet transform has been used extensively in several areas of signal processing applications such as signal prediction, speech processing, biomedical engineering, image denoising, image annotation, image/video watermarking and video processing.

A multi-resolution representation of a signal is achieved by the wavelet transform using a set of orthonormal analysis functions produced by some base functions called ‘wavelets’ [34, 39]. One of the major attributes of this transform is its ability to characterize the spatio-temporal coherence between the signal components that is of concern in our analysis. 3D wavelets present a division of video spectrum into multi-scale subbands for temporal and spatial dimensions; and oriented subbands for spatial information –the horizontal, vertical and diagonal subbands. This transform is separable and the decomposition is done by passing through a 3D-filter channel bank. Each 3D-filter channel bank can be identified as a multiplication of three one-dimensional filter banks.

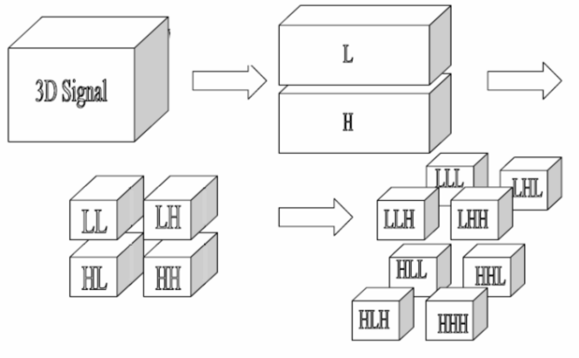


Figure 1. Implementation of one-level 3D wavelet transform reproduced from [40].

At each level of decomposition, eight subbands are created. The total number of coefficients of all subbands is the same as the size of the original signal, where each time the lowest subband, LLL, contains an approximation to the original signal. The next level of decomposition is done merely within the lowest subband [41]. The implementation of the one-level 3D wavelet transform is shown in figure 1, where L and H stand for Low and High subbands, respectively.

3.1. 3D Wavelet Coefficients Relationships

Each 3D wavelet subband stands for a sub-sampled version of the filtered original signal; hence, there are some relationships between coefficients in different subbands corresponding to the same part of the original signal [27, 34, 62]. Coefficients in different subbands with the same orientation have parent-child relationships. Coefficient at the same level and location with different orientations are cousins, whereas adjacent coefficients in the same subband are neighbors.

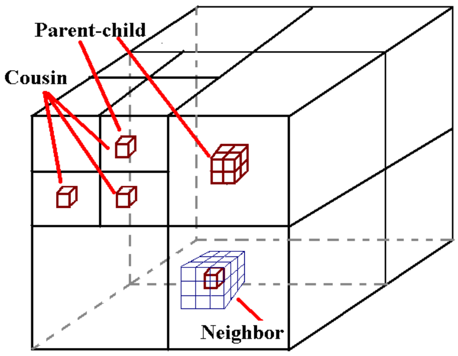


Figure 2. 3D wavelet transform coefficients relationships, deduced from [27].

The 3D wavelet coefficients relationships are shown in figure 2. Each coefficient X in a finer scale subband has 6 cousins, named CX, and 26 neighbors, named NX. Each coefficient in a coarser scale subband has 8 children in the corresponding finer scale subband; thus each coefficient X in a finer scale subband has a parent (PX) in the coarser scale subband. Consider the wavelet coefficient $w_{l,\boldsymbol{P}}^{o}$ at the orientation o, where $o = 1,2,\dots,7$ for orientations $LLH, LHL, LHH, HLL, HLH, HHL, HHH$ respectively, and at the transform level l, where $l = 1,\dots,lev$, and location $\boldsymbol{P}$=(x,y,t). So, the coefficients in the same scale and position and different orientations, $\{w_{l,\boldsymbol{P}}^{1}, w_{l,\boldsymbol{P}}^{2}, \dots, w_{l,\boldsymbol{P}}^{7}\}$, are cousins. Also, $\{w_{l,(x+i,y+j,t+k)}^{o}, i = -1,0,+1, j = -1,0,+1, k = -1,0,+1, (i,j,k) \cong \boldsymbol{0}\}$ represents the neighbors of coefficient $\boldsymbol{P}$=(x,y,t) in the orientation o and transform level l. Finally, the wavelet coefficient $w_{l,\boldsymbol{P}}^{o}$ is the parent of coefficients $\{w_{l+1,(2x+i,2y+j,2t+k)}^{o}, i = 0,1, j = 0,1, k = 0,1\}$, in the finer scale and the same orientation.

4. 3D Wavelet Statistics

4.1. Background

A 2D wavelet based image modeling was introduced in [42] based on probability modeling. Experimental results show that the Generalized Gaussian Density (GGD) function could yield a suitable estimate of the density of the 2D wavelet coefficients of each subband, using different filters [27, 33, 43-45]. Minh N. Do used the GGD features and the Kullback-Leibler Distances (KLD) for texture retrieval and gained good results [27]. So, two extracted GGD parameters, α and β, will help to estimate the density function of the coefficients.

The approximation to the probability distribution function (PDF) for the marginal density of a signal can be achieved by adaptively changing the two parameters of the GGD [46], defined as:

$$p(x;\alpha,\beta) = \frac{\beta}{2\alpha\Gamma(\frac{1}{\beta})} e^{-(\frac{|x|}{\alpha})^{\beta}} \tag{1}$$

where $\Gamma(.)$ is the Gamma function.

$$\Gamma(z)=\int_0^{\infty} e^{-t}t^{z-1}dt, \qquad z>0 \tag{2}$$

In (1), α is the scale parameter that models the width of the GGD and β is the shape parameter proportional to the inverse of the decreasing rate of the PDF. The PDF is Laplacian if $\beta=1$, and Gaussian if $\beta=2$.

4.2. Dataset for studying the statistics

To study the statistical characteristics of the 3D wavelet transform and to investigate the activity level in the scene, we have conducted experiments using various natural video signals of different types of activity and texture from Hollywood-2 Human Actions and Scenes dataset (CVPR09) [47], the 'Simon Fraser University (SFU) Video Library and Tools dataset' [48], and 'The Open Video Project database' [49]. We have randomly selected 750 video samples from the sequences in these datasets to evaluate our method that is based on the statistical properties of the 3D wavelet video. Selected video clips, consisting of more than 120,000 video frames of different characteristics, are employed to evaluate our ideas. The above video repositories are widely used in the community. The spatial size of the videos is different including CIF (352x288) or QCIF (176x144), and their temporal length is between 75 to 300 frames, where the sampling rates are 15, 24 and 29 frames per second. Each video sample contains a single video shot. Video tests are captured from different locations – indoor/outdoor, as well as home, road, library, store, office, etc. They display human actions, car racing, news broadcasting, dogs running, airplane flying, glass breaking, etc. Also, they contain camera zooming, panning, translations and scene fade in/outs. Most of the test videos are highly dynamic in both temporal and spatial domains, though containing a few static samples.

We have also used well-known filters – Haar, Daubechies and Symlets - to decompose the signals. Three, four or five decomposition levels have been selected in the tests. To classify activity levels, we have considered two domains for activities:

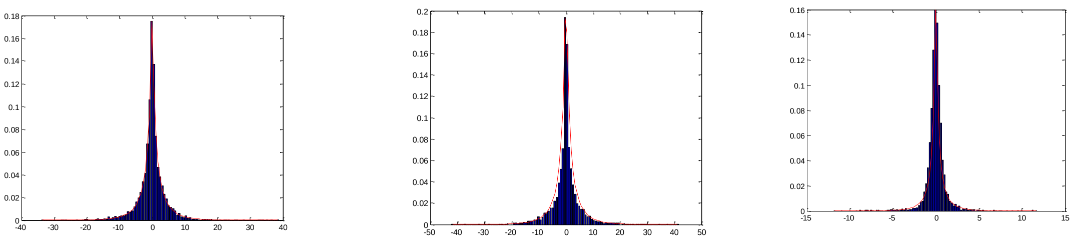
- **Temporal domain:** This considers the speed of the changes in the time domain and is taken to be slow or fast.

- **Spatial domain:** This considers the spatial fraction of changing pixels involved in an event or activity. It can be local or global.

We then obtained the 3D wavelet statistical properties of the test videos and analyzed them, as discussed next in 4.3 and 4.4.

4.3. Marginal statistics

We have applied the 3D wavelet transform with various filters to our video test set, and studied the marginal statistics of the resulting 3D wavelet transforms. In particular, we have applied the transform to the 750 video sequences mentioned in section 4.2, and extracted the GGD parameters from all of the subbands and checked the marginal histograms with the GGD curves. It was found from this experiment that 100% of the curves fitted perfectly to the histograms, confirming that the GGD could give a close estimate of the signal histograms. Also kurtosis values of these curves - as a measure of the 'peakedness' of the probability distribution of real-time random variable, are calculated from these more than 21000 subbands, which are all larger than 9, stating non-Gaussian quality of these distributions. Three normalized histograms of three different signals and the fitted curves are shown in figure 3. As shown in the figure, each histogram has a peak at zero and decreases rapidly to zero. This means that most of the coefficients are zero or near zero; thus, it confirms that the 3D wavelet transform is very sparse. The values of α , β and kurtosis are also given in the figure. The kurtosis values of these three histograms are 15.6221, 11.8763 and 25.7525 which prove the highly non-Gaussian property of the densities, considering that the kurtosis value of a Gaussian density is 3. So, estimating two GGD parameters from each video subband could give sufficient information about the marginal quality of the associated subband, and employing two GGD parameters, the marginal distribution of wavelet coefficients in a subband can be captured precisely, where hundreds of features are required to indicate this information using histograms.



$\alpha = .7588, \beta = .5819,$
$Kurtosis = 15.6221$
$LHH - Subband$

$\alpha = .8591, \beta = .5788,$
$Kurtosis = 11.8763$
$HHH - Subband$

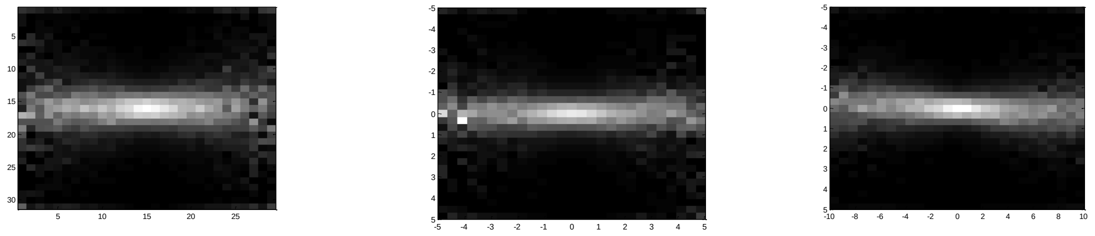
$\alpha = .1203, \beta = .5002,$
$Kurtosis = 25.7525$
$HHL - Subband$

Figure 3. Marginal histograms of the finest 3D Wavelet subbands and GGD curves fitted to them.

4.4. Joint statistics

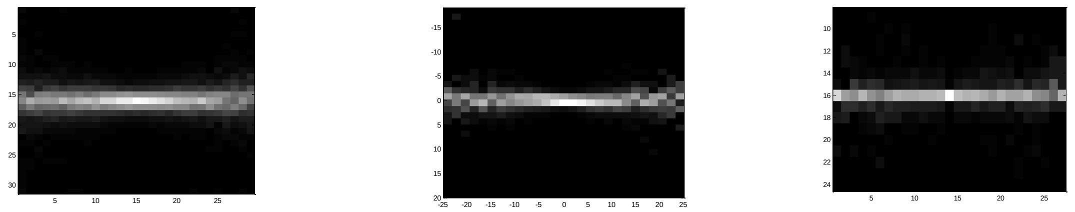
The joint statistics of the 2D wavelet and the 2D contourlet transforms have been extracted in [42, 46], respectively. Here, we have worked on the joint statistics of the 3D wavelets of our video test set. Although the 3D wavelet transform finely decorrelates the video signal, there are still dependencies among coefficients of different subbands in the same scale and of the same subbands in different scales. Video processing algorithms can be developed based on the joint statistics of the coefficients. One of the joint statistics plots for the 3D wavelet coefficients, conditioned on their parents, neighbors and cousins, is shown in figure 4.

The conditional plots have the form of "bow–tie" where their variance and magnitude are interrelated [50]. Furthermore, the conditional expectations are about zero; therefore, coefficients are dependent and almost uncorrelated. The joint density plots are also presented in figure 5 conditioned on two distant parents, neighbors, and cousins. Results confirm the independencies between these coefficients; i.e., the dependencies between coefficients and their parents, neighbors, and cousins are local and decrease sharply when the distance increases. One of the vertical cross sections of the joint statistics plots is shown in figure 6. The kurtosis values of these histograms specify that the conditioned distributions are still non-Gaussian.



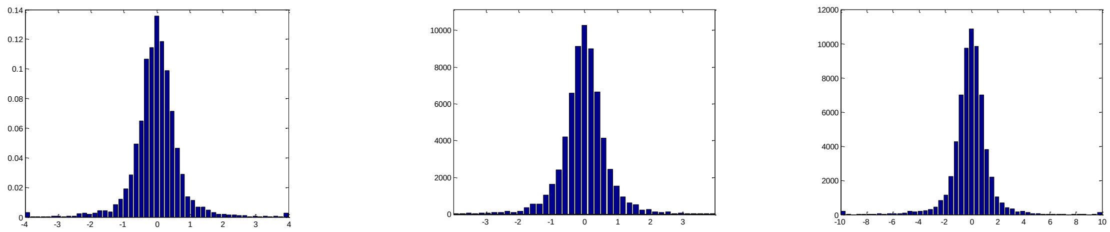
(a) (b) (c)

Figure 4. Conditional plots of coefficients conditioned on their (a) parents, (b) neighbors, (c) cousins. The plot is sketched for only one of the neighbors (here the right side neighbor in spatial X direction) and one of the cousins (here the cousin in the HLH for the coefficient in HHH subband).



(a) (b) (c)

Figure 5. Conditional plots of coefficients conditioned on their distant (a) parents, (b) neighbors and (c) cousins. The plot is sketched for only one of the neighbors (here the right side neighbor in spatial X direction) and one of the cousins (here the cousin in the HLH for the coefficient in HHH subband).



(a) kurtosis=5.8791 **(b) kurtosis= 4.0841** **(c) kurtosis=7.8338**

Figure 6. Vertical cross section of joint statistics plots (a) parents, (b) neighbors, (c) cousins.

5. Mutual Information Estimates and Video Activity Analysis

5.1. Background

In this section, we use the MI (Mutual Information) as a suitable quantitative dependency measure. Although correlation is a good meter of dependency in Gaussian distribution, it is not applicable to non-Gaussian cases [51], like our case and hence we use the MI. The MI between two continuous random variables *X* and *Y* is computed as [46, 52]:

$$\begin{aligned} I(X;Y) &= \int_x \int_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dydx \\ &= E_{XY}\{\frac{p(x,y)}{p(x)p(y)}\} = D(p(x,y) \| p(x)p(y)) \end{aligned} \tag{3}$$

where $p(x,y)$ is the joint density function between *X* and *Y* and $p(x)$ and $p(y)$ are the marginal PDFs of *X* and *Y*, correspondingly. $E\{\}$ stands for the expected value and $D(.)$ represents the Kullback-Leibler Distance (KLD). This value shows the amount of information of one variable in relation to the other. We use base 2 logarithm; hence, $I(X;Y)$ is measured in bits. The MI is the amount of information variable *X* transmits about variable *Y*, and vice versa, thus the MI is non-negative and symmetric. Moreover, the MI shows the amount of dependency between two variables and will be zero if *X* and *Y* are independent. On the other hand, the MI increases by increasing the variable dependency.

To estimate the MI, the distributions histograms are used in the following formula [46, 53]:

$$\hat{I}(X;Y) = \sum_{i,j} \frac{h_{ij}}{N} \log \frac{h_{ij} N}{h_i h_j} - \frac{(J-1)(K-1)}{2N} \tag{4}$$

where h_{ij} is the value of the cell (i,j) in the joint histogram, $h_i = \sum_j h_{ij}$ and $h_j = \sum_i h_{ij}$ are the marginal histograms, *N* is the number of all coefficients, and *J* and *K* are the number of bits in *X* and *Y* directions, respectively. The second term in (4) is a modification partial bias and cannot resolve the problem entirely. This expression tries to reduce the error, but cannot remove all of it; thus, equation

(4) just introduces a lower bound for the MI [46, 53]. In [46] the values of J and K are chosen experimentally as shown below to make the estimation firmer, while the error in equation (4) increases when the number of variables increases:

$$J = K = round(\frac{N}{3000}) + 1 \tag{5}$$

5.2. Results and Analysis

We used the MI estimates to study dependency between parent, child, neighbors, and cousins in the 3D wavelet transform of natural videos, and then the relationship between activity levels in the video and the amount of the MI were interpreted. The video test set used here is the same as the set employed in the previous section. To do so, the estimates are studied in three steps. In the first step, the MI estimates between a coefficient of the finest level and its parent, neighbor and cousin are computed and some of the results are illustrated in table 1 and figure 7. In this table the MI estimates between the coefficient and its 6 cousins and MI between the coefficient and its 26 neighbors are calculated and averaged.

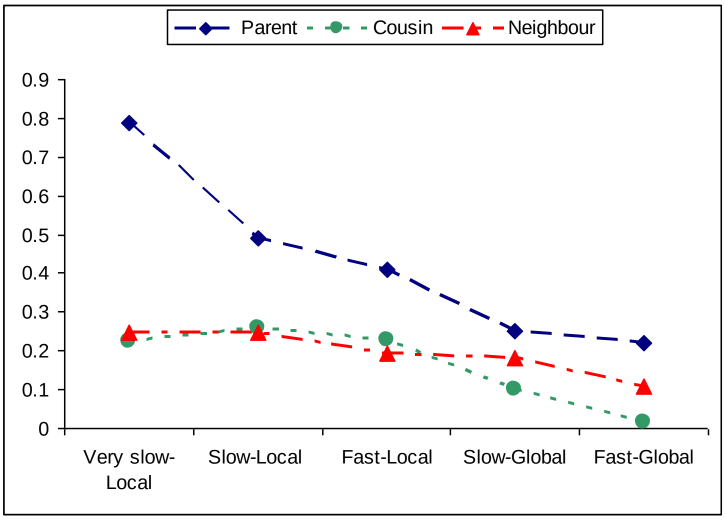


Figure 7. MI estimates between X and its parent (PX), neighbor (NX), cousin (CX), 3-level Wavelet ('db4').

Table 1. MI estimates between X and its parent (PX), neighbor (NX), cousin (CX), 3-level Wavelet ('Daubechies filter').

	Very slow-Local	Slow-Local	Fast-Local	Slow-Global	Fast-Global
$I(X;PX)$	0.7863	0.4887	0.41	0.2516	0.2217
$I(X;CX)$	0.2254	0.259	0.2277	0.1003	0.0152
$I(X;NX)$	0.2363	0.2467	0.1927	0.1825	0.1096

The results can be used to infer the following:

- As it is deduced from the table 1 and the figure 7, the MI is significantly high, which confirms the results given in subsections 4.4., where the results, obtained from the joint statistics histograms, showed the dependency between the coefficients and their parents, cousins or neighbors.

- The MI between the coefficient and its cousin as well as the MI between the coefficient and its neighbor has the least value for the high activity level video. It confirms that the dependency between different subbands in the same level is small.

- The MI between the coefficient and its parent always has the highest value. It means that the main dependency is between the parent and its child, where increasing the decomposition level, more delicate information is extracted.

- The MI increases when the activity in the video decreases. Thus the dependency between the coefficients and their parents, cousins and neighbors decreases by increasing the activity level in the video.

- The changes in the MI between the coefficient and its neighbor based on activity changes are smaller than the changes in the MI between coefficient and its parent or cousin.

In the second step, the mutual estimates are calculated for different types of wavelet filters (table 2). It is inferred from the results that the MI is dependent on the filter type. For example, replacing the "Haar" by the "Daubechies" reduces the mutual estimates and dependencies between the coefficients.

Table 2. MI estimates between X and its parent (PX), neighbor (NX), cousin (CX), Different filters (high activity video, 3-level).

	Haar	Daubechies	Symlet
$I(X;PX)$	0.2454	0.2217	0.1233
$I(X;CX)$	0.0238	0.0152	0.0168
$I(X;NX)$	0.2023	0.1096	0.1645

In the third and final step, the MI is estimated for different numbers of transform levels (table 3). As shown, the estimates between the coefficient and its cousin/neighbor are not much dependent on the number of transform levels.

Table 3. MI estimates between X and its parent (PX), neighbor (NX), cousin (CX), Different levels (high activity video, 'Daubicies filter).

Transform Levels →	2	3	4
$I(X;PX)$	0.1960	0.2217	0.2797
$I(X;CX)$	0.0152	0.0152	0.0152
$I(X;NX)$	0.1096	0.1096	0.1096

6. Activity Analysis based on 3D Wavelet statistics

6.1. Joint Statistics and the Kurtosis Curves

Here, we have used the video samples described in section 4.2 and have applied the 3D wavelet to all of the test clips. Temporal and spatial fields are separately observed for the activity level analysis. The changes are categorized into slow or fast, depending on the scene change rate, and also identified as local or global, due to their relative surface within a given frame.

The joint statistics of the 3D wavelet transform was shown in subsection 4.4. The distributions conditioned on the parents of the finest level are used here to classify videos according to their activity levels. To produce these curves, the distributions conditioned on the parents of the finest level are calculated and the kurtosis values of vertical cross section histograms of these distributions are computed. These kurtosis values are computed from the kurtosis curve of each conditioned distribution. These curves are calculated for all seven finest subbands of wavelet transforms of each video sample and the corresponding elements of these seven curves are accumulated and averaged to form the final kurtosis curve. The kurtosis curves of the vertical cross section histograms of coefficients, conditioned on parents, have been extracted from videos of different types. Some of these curves are depicted in figure 8. To produce these curves, the distributions conditioned on the parents of the finest level are calculated and the kurtosis values of vertical cross section histograms of these distributions are computed. These kurtosis values are used to from the kurtosis curve of each conditioned distribution. These curves are calculated for all fines subbands of each video sample and the corresponding elements of these seven curves are accumulated and averaged to form the final kurtosis curve. As shown, there are four major types of curves, each matching with a level of activity in the video clip. Accordingly, we can cluster videos based on their kurtosis curves into four groups:

- **Group 1: Very high activity level videos -** videos with an object emergence, fast global movement, very fast changes or noisy videos. The kurtosis curve in this group is smooth and beneath 5 at zero. Thus, the shape of the curve is nearly flat. There is no apparent peak at zero and the kurtosis decreases slowly by increasing the absolute value of the parent coefficient until arriving at 3.

- **Group 2: High activity level videos -** videos with slow-global movements. The kurtosis curve in this group is beneath 10 and above 5 at zero. There is a peak at zero and the kurtosis decreases gradually by growing the absolute value of the parent coefficient until reaching 3.

- **Group 3: Low activity level videos -** videos with slow-local movements. Here, the kurtosis curve is between 30 and 40 at zero. There is a peak at zero and the kurtosis decreases rapidly by increasing the absolute value of the parent coefficient until it reaches 3 (Gaussian), thus the curve is sharp around zero and almost flat thereafter.

- **Group 4: Very low activity videos -** videos with very slow and local changes. Shape of the curves in this group resembles shape of the curves in the third group, where it is too sharp at zero. The kurtosis curve has a peak around 40 at zero and decreases rapidly by increasing the absolute value of the parent coefficient, thus the curve are very sharp.

The curves, shown in figure 8 demonstrate that by increasing the activity in the videos, the kurtosis value decreases and the coefficients distributions, conditioned on parents, get closer to Gaussian. As discussed in subsection 5.2, the dependency between the child and parent decreases by increasing the activity in the video. Thus, the kurtosis value decreases by decreasing the dependency and increasing the activity.

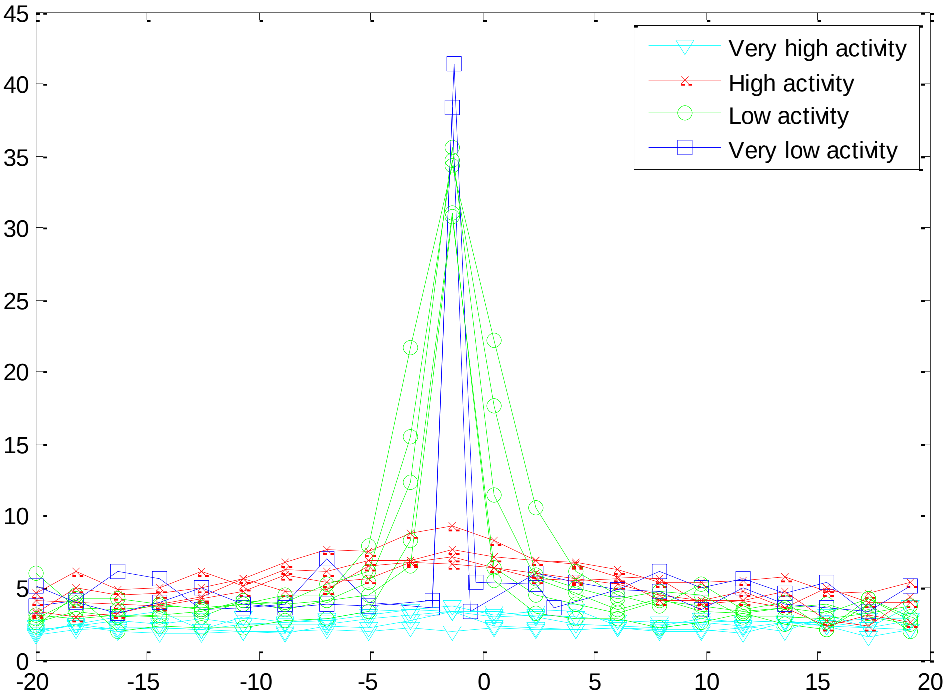
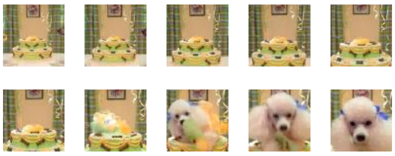


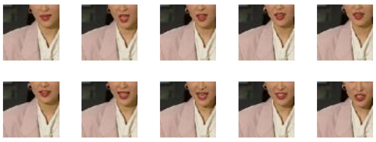
Figure 8. Four different kurtosis curves.

To do the classification, the distributions conditioned on the parents, neighbors and cousins of all the seven subbands of the finest level are calculated and the kurtosis values of the vertical cross section histograms of these distributions are employed to form the feature vector. We have used nine bins for each curve and the kurtosis curves, conditioned on parents, cousins and neighbors are calculated for each video sample as described. Thus each video will have three 9-bin kurtosis curves and 27 joint features as a result. We use the SVMs [54] to classify the kurtosis curve features into four classes. For this purpose, 500 video samples are employed in the training sets that are labelled manually. We have extracted kurtosis curves values as the feature vectors of 58 different video signals and categorized them into four groups utilizing the trained classifier. A sample video sequence of each group is shown in figure 9.

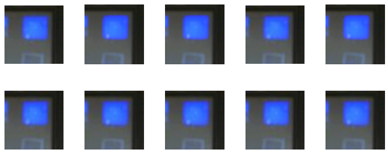


(a) Object emergence – group 1

(b) High textured – group 2



(c) Slow local scene – group 3



(d) Very low activity – group 4

Figure 9. Some Sample videos and their identified groups; the frames are downsampled at rate 3.

We have evaluated the performance of the proposed grouping algorithm by conducting subjective tests. Fifteen non-expert observers are employed to group the video samples into four activity levels, according to ITU-R recommendation BT.500-11 [61]. Testers are asked to rate the videos based on their activity level. The subjective test is a single stimulus test [55], where questionnaires are designed and each individual has been asked to fill the table once each video sample is played. The subjects had normal visual perception, having no special knowledge about video processing methods. The instructions along with some illustrations were given to the individuals, prior to the test. First, the video sequence was played in the top part of the screen, then each subject was asked to rate the activity level in the video sequence by an integer number between 1 and 4 –for activities from highest to lowest, respectively. The average results of the subjective tests are shown in figure 10.

We have also compared our method to a baseline method, in which the sum and energy of the coefficients of wavelet transform of each video subband are used as in [36] to form the feature vectors of the video samples. The 3D wavelet transform based features of salient regions as well as geometric based features are employed in [36] for human action recognition, where a simple leave-one-out approach is used to evaluate the results. Here again SVM classifier is applied and the training and test stages are the same as the above mentioned procedure for the proposed method.

Based on subjective test results, our method is capable of grouping the video sequences with 87.3% accuracy, as compared to the classification performed by humans. Also comparing this result with the baseline method, it outperforms this method too. This shows that the joint statistics convey important and key information about the speed and amount of changes in the video sequence. The grouping rates corresponding to each group are also sketched in figure 10.

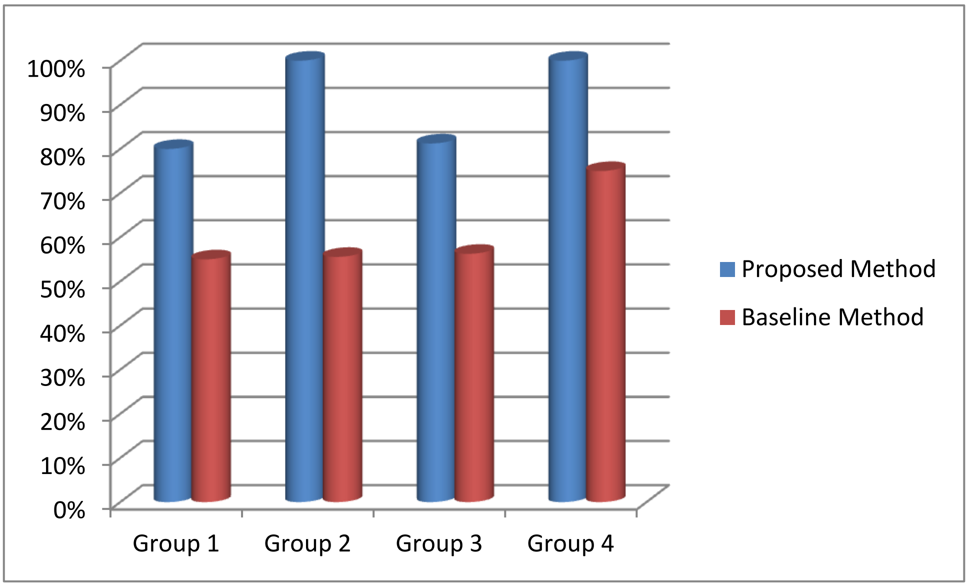


Figure 10. Grouping accuracy rates for each class.

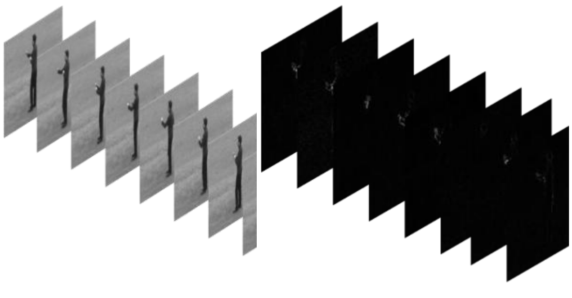
6.2. Marginal and Joint Statistics applied to Human Action Recognition

We employed the 3Dwavelet transform features for human action recognition. The KTH human action database is used to evaluate our method [18]. This database consists of 2391 video sequences performed by 25 persons in 4 different scenarios of 6 different actions including boxing, hand clapping, hand waving, running, jogging and walking. The spatial size of the frames is 160x120 pixels and the durations of the sequences are different and the temporal sampling rate is 25 frames per second. The videos are taken in four scenarios: outdoors, outdoors with scale variations, outdoors with different clothes, and indoors. The camera view point varies in different video samples, but the camera is mostly static. This dataset is divided into training (8 subjects), validation (8 subjects) and testing (9 subjects) sets based on the performers. Figure 11 shows some sample frames of 6 different actions (columns) taken in 4 scenarios (rows). The KTH database is employed by several human action recognition algorithms to evaluate their performances [10-12, 16-18, 36], where [10, 16, 36] have employed simpler tests and are therefore left out in our comparisons.

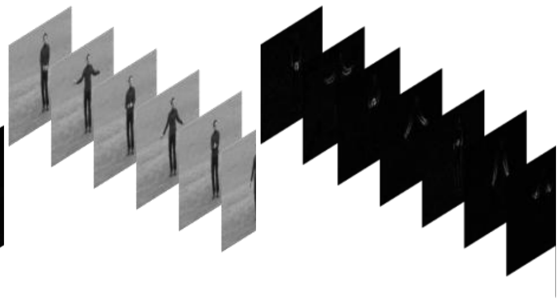
Figure 11. Sample frames from KTH database [18].

6.2.1. The Proposed Algorithm

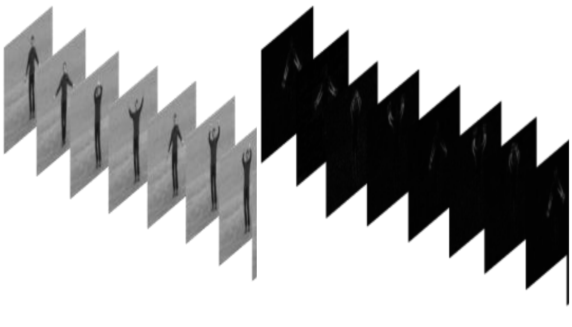
Our method can be considered as a global image representation approach to human action recognition. The general concept in this approach is to remove the background and extract features from the remaining human silhouette, which has the information of the motion and the shape of the body. The background in the KTH database is not static and background extraction is not a simple step. Due to the characteristics of the 3D wavelet transform; i.e., its sensitivity to the edges and their variations over time, we have used the differences of adjacent frames instead of the extracted human body as in [5, 64]. The resulting subtracted video sequence contains the desired motion information. Next, we have applied 3D wavelet transform to the subtracted video sequence, extracted marginal and joint statistics parameters from each video sample as described in section 4, and constructed the feature vectors. Finally, Support Vector Machines (SVM) are employed to classify the feature vectors. We have used the LIBSVM [54] library to train, validate and test the accuracy of the classifiers.



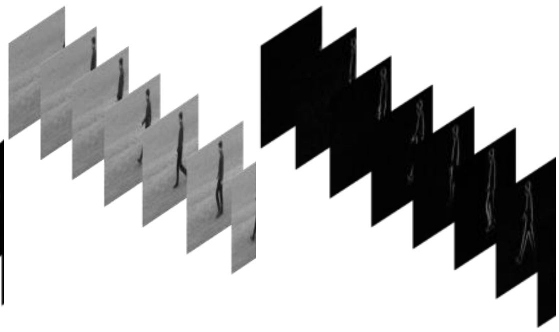
(a) Boxing



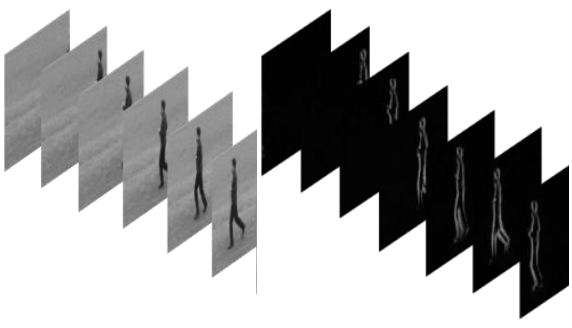
(b) Hand Clapping



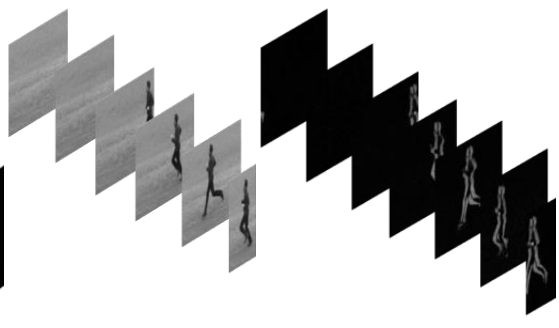
(c) Hand Waving



(d) Walking



(e) Running



(f) Jogging

Figure 12. Downsampled video sequences (left) and the subtracted video sequences (right).

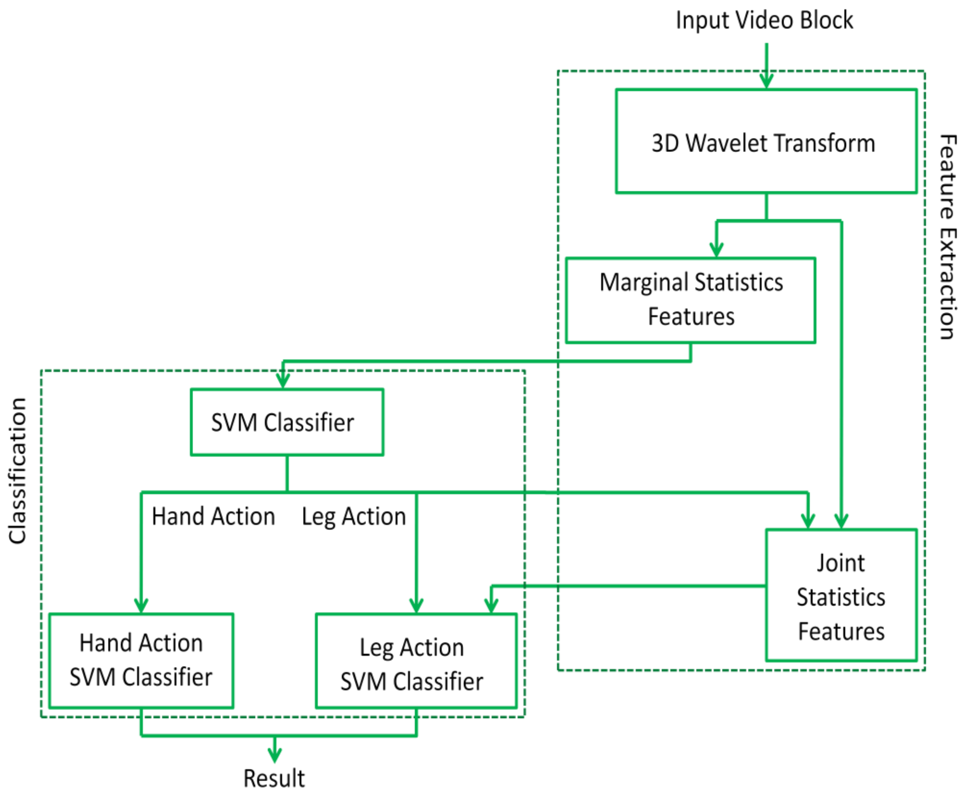


Figure 13. the proposed algorithm

As deduced from our experiments, the 3D wavelet transform coefficients plausibly convey the information about the activity in the video sequences. This information consists of the rate and the direction of movements. Figure 12 shows sample video sequences and their corresponding subtracted video sequences per each action class. It is observable in the figures that the type of edges and movements in the hand actions – boxing, hand waving and hand clapping - is completely different from the leg actions – walking, running and jogging. For hand actions recognition, the types of movements are the leading elements, whereas the rates of motions are the important factors to classify leg actions. Based on the above analysis, a hierarchical classification was applied to human action recognition. First, we classify the actions into two classes: legs and hands, based on the marginal statistics features of the video samples employing the SVM classifier. In the next step, two different classifiers are used to classify the hand and leg actions separately. For leg actions the joint statistics features, as well as marginal statistics features, are utilized for classification, while only marginal statistics are used for hand actions.

The proposed algorithm is illustrated in figure 13. First, the 3D wavelet transform is applied to the subtracted video sequence. Second, marginal statistical parameters, the GGD features, are extracted from each video subband based on the coefficient distributions of the wavelet subbands employing equation (1). The extracted GGD features are used to construct the marginal feature vector of each video sequence. The number of extracted marginal features are $14\times lev$, where there are two features – α and β- extracted from each video subband and seven video subbands at each level, and lev is the number of transform levels. Then, these marginal features are applied to the first SVM classifier to classify the 'leg' and 'hand' actions. If the action is classified as a 'hand' action, the marginal feature vector goes through the 'Hand action SVM classifier' to recognize the hand action.

In the case of 'leg' action classification, joint features should be extracted from the video sequence. The distributions conditioned on the parents, neighbors and cousins of all the 7 subbands of the finest level are calculated and the kurtosis values of the vertical cross section histograms of these distributions are calculated and the final joint feature vector consisting 27 features is constructed- as explained in section 6.1- for the video sample. Also, the MI between the coefficients and their parents of the seven subbands of the finest level are estimated and used to complete the joint feature vector of length 34. Finally, these features along with the marginal features are employed to detect the 'leg actions' using the 'Leg action SVM classifier'.

6.2.2. Human Action Recognition Results

We have computed the recognition rates for different numbers of the wavelet transform levels and different wavelet filters in tables 4(a) and 4(b), respectively. Results show that the best result is gained by the wavelet filter "Symlet" where increasing the transform level increases the recognition rate, except for the 5-level transform. The reason is that increasing the wavelet transform levels will lead to more edge and motion details being considered by the features and in the KTH database with non-static background, this causes interference with the necessary and effective data, leading to undesired information from background noise. So, this causes the classification rate to decrease. Also, by increasing the levels, more noisy features might be employed by the SVM classifier that may reduce the classification accuracy. Recognition accuracies of our best classification result and other reported methods are compared in table 5 for KTH Dataset.

As mentioned earlier, the methods given in [10, 16, 36] have employed leave-one-out tests, which are simpler than our classification method, where the training set is fully used to train the classifier and all the test videos are used together to evaluate the classifier. The highest recognition rates reported in [10], [16] and [36] are 91.7%, 91.6% and 58.9%, respectively, while we have achieved an accuracy rate of 94.6%, in the leave-one-out test procedure. In this test, the ‘Symlet’ wavelet filter and four transform levels are used to decompose the video sequence and extract spatio-temporal features.

Table 4. Average Action Recognition Rates for KTH Dataset.

(a) 3D Wavelet transform, “Symlet” wavelet filter, different transform levels.

Number of Transform Levels	Precision
2 levels	82.1%
3 levels	89.7%
4 levels	**93.4%**
5 levels	91.8%

(b) 3D Wavelet transform with 4 levels, different wavelet filters.

Number of Transform Levels	Precision
Haar	88.3%
Daubechies	90.8%
Symlet	**93.4%**

Table 5. Comparison of different Human Action Recognition methods for KTH Dataset.

Technique	Method in [18]	Method in [12]	Method in [17]	Method in [11]	Our proposed method
Precision	71.7%	83.3%	86.7%	91.8%	**93.4%**

6.2.3. Discussion

In this subsection, the human action recognition results are discussed in detail to state the reasons of its efficiency. Also, the computational complexity of the proposed method is investigated.

The confusion matrices of our method (93.4%) and the same approach when no joint statistics features are used in the classification (92.5%) are shown in tables 6.a and 6.b, respectively. Also the confusion matrix of the method [11], the best of the existing methods on KTH database for the action classification, is depicted in table 6.c. Comparing these three matrices, it can be inferred that the hand

and leg actions are well differentiated and the main confusion always occurs in the leg action classification, especially between 'jogging' and 'running', where our method does a better job for that specific classification. The reason is the ability of the wavelet transform to locate edges and their movements, textures in the video, as well as the details that are contextually more important to the human visual system [27, 34]. The important factor to differentiate between three 'leg' actions is the speed of changes in the video sequence, which is well addressed by the wavelet transform features, as was expected based on the analysis given in sections 4.4 and 5.2 about joint statistics and mutual information between each coefficient and its parent, cousin and neighbor. Moreover, our method uses global features instead of local ones which make it simpler in implementation and classification.

Table 6. Comparison of the confusion matrices of the proposed method (tables 6.a and 6.b) and the method in [11] (table 6.c) for KTH Dataset.

6.a. With joint statistics features

	walk	Jog	Run	Box	Hclp	Hwav
Walk	100	0	0	0	0	0
Jog	2.1	91.7	6.2	0	0	0
Run	0	15.3	84.7	0	0	0
Box	0	0	0	99.3	0	.7
Hwav	0	0	0	2.8	93.75	3.5
Hclp	0	.7	0	8.3	0	91

6.b. Without joint statistics features

	walk	Jog	Run	Box	Hclp	Hwav
Walk	100	0	0	0	0	0
Jog	2.1	91	6.9	0	0	0
Run	0	20.1	79.9	0	0	0
Box	0	0	0	99.3	0	.7
Hwav	0	0	0	2.8	93.75	3.5
Hclp	0	.7	0	8.3	0	91

7.c. The method in [11]

	walk	Jog	Run	Box	Hclp	Hwav
Walk	99	1	0	0	0	0
Jog	4	89	7	0	0	0
Run	0	19	80	0	0	0
Box	0	0	0	97	0	3
Hwav	0	0	0	0	91	9
Hclp	0	.7	0	5	0	95

To discuss the complexity of the proposed method, we first consider that the computational complexity of the 1D discrete wavelets transform for an N-length vector matrix is $O(N)$ [57]. Thus, we can assume that a positive non-decreasing linear function $f_1(N)$ can be found to represent the complexity of the wavelet transform. Since the 3D wavelet transform is applied to each dimension of the signal separately, for the video block of the spatial size of $X \times Y$ and the temporal size of T frames, it can be assumed to have XY, XT and YT 1D wavelet transformations on the vectors with T, Y and X lengths, respectively. The complexity of the 3D wavelet transform will therefore become $XYf_1(T) + XTf_1(Y) + YTf_1(X)$. To be more precise, for each level of 1D wavelet transform of a vector of length N, there will be $N \times l$ multiplications, where N is the number of coefficients and l is the length of the wavelet filter. So the total number of multiplications will be $N \times l + \frac{N \times l}{2} + \cdots + \frac{N \times l}{2^{lev-1}} \leq N \times l \left(1 + \frac{1}{2} + \cdots + \frac{1}{2^{lev-1}} + \cdots + \frac{1}{2^n}\right) = 2N \times l$, where $n \to \infty$. Consequently, the overall computational complexity of the 3D wavelet transform for a matrix of size XYT will be upper-bounded by $6lev(XYT)$, or by $6l(TX^2)$, when $X = Y$.

The GGD parameter estimation algorithm has computational complexity of $O(N)$, where N is the number of samples [27]. Thus, the order of calculations can be represented by a positive non-decreasing linear function $f_2(.)$ - f_2 is big-O of N. The number of samples in the subbands of the first decomposition level is $\frac{XYT}{8}$ and this amount is divided by 8^{lev} for other levels. Since we have seven subbands at each level, we will have $7.\frac{XYT}{8}(1 + \frac{1}{8} + \cdots + \frac{1}{8^{lev-1}})$ which is a geometric progression equal to $7.\frac{\frac{XYT}{8}\left(1-\frac{1}{8^{lev}}\right)}{1-\frac{1}{8}} < 7.\frac{\frac{XYT}{8}}{1-\frac{1}{8}}$, where the right side of the inequality equals to XYT. Thus, the complexity will be the same as that of the 3D wavelet transformation stage; it is the same as what is claimed in [58] about the time complexity of the GGD parameter estimation. Also, the computational complexity of the calculation of the joint parameters and the corresponding kurtosis value is $O(XYT)$. Again, there will be a positive non-decreasing linear function $f_3(.)$ - f_3 is big-O of N- to stand for the computational complexity of this stage. So the overall complexity of feature extraction step will be less than $f_2(TXY) + f_3(TXY)$, where $X = Y$. Accordingly, a positive non-decreasing linear function $g_2(.)$ can be assumed, such that the inequality of $f_2(TXY) + f_3(TXY) \leq g_2(TX^2)$ is kept true. This computational complexity is of the same order as that of the l_2-norm feature extraction used in [36], while the results are much better. Considering the same assumption about positive non-decreasing linear function $g(.)$ - $6l(TX^2) + g_2(TX^2) \leq g(TX^2)$- the overall complexity of the feature extraction algorithm can be represented by $g(TX^2)$.

For the signal classification, as described earlier, a SVM classifier is employed, with the computational complexity in the order of $O(kn_{fv}{n_{tr}}^2)$ for the training phase and $O(kn_{fv}n_{tr})$ for the test phase, where k is the number of classes, n_{fv} is the number of features and n_{tr} is the number of the training samples [59,60]. In our algorithm, there are $14lev$ features extracted from marginal statistics, where lev is the number of transform levels which is always 4, and 34 features from joint statistics investigations. For the recognition phase, we will consider two classification scenarios:

First, the SVM classifier is used to classify all six actions into six classes. Here, n_{fv}=90, $n_{tr} = 1536$ and $k = 6$, so the complexity can be represented by positive non-decreasing linear functions $h_1(k, n_{fv}, n_{tr}^2)$ for the training phase and $h_2(k, n_{fv}, n_{tr})$ for the test set. In the second scenario, the proposed hierarchical SVM classification scheme is employed, where first all actions are taken to classify the leg and hand actions, using the marginal features. Then, hand actions are classified again using the GGD features, where the number of training samples is halved and number of classes is 3. For the leg actions, the number of features is increased by the number of joint features. So, the complexity of the training step can be represented as:

$$h_1\left(\frac{k}{3}, \frac{56}{90}n_{fv}, n_{tr}^2\right) + h_1\left(\frac{k}{2}, \frac{56}{90}n_{fv}, \left(\frac{n_{tr}}{2}\right)^2\right) + h_1\left(\frac{k}{2}, n_{fv}, \left(\frac{n_{tr}}{2}\right)^2\right) = 0.413 \times h_1\left(k, n_{fv}, n_{tr}^2\right) \qquad (6)$$

which is less than half of the complexity when a single SVM is used. Also, the complexity of the test phase is given as:

$$h_2\left(\frac{k}{3}, \frac{56}{90}n_{fv}, n_{tr}\right) + h_2\left(\frac{k}{2}, \frac{56}{90}n_{fv}, \frac{n_{tr}}{2}\right) + h_2\left(\frac{k}{2}, n_{fv}, \frac{n_{tr}}{2}\right) = 0.62 \times h_2(k, n_{fv}, n_{tr}) \qquad (7)$$

which is less than the complexity of using a single SVM classifier.

These calculations show that applying the hierarchical SVM to the human action recognition task has decreased the complexity of the method.

To compare the complexity of the proposed method with that of the method introduced in [11], we first observe the classification stage, where both methods have employed SVM classifiers. The number of features used for the classification linearly affects the computational complexity of the system. In [11] 4000 words have been selected empirically to produce the feature vectors, where in our method, in the case of using one SVM classifier that is the worst case, we have employed about 100 features; thus, the complexity of the proposed method will be alomost 0.1 less than the method in [11].

We ignore comparing the feature extraction complexities, since the method in [11] employs interest points for features extraction. To select these points, a spatio-temporal scale-space representation should be calculated applying the convolution of the video block with Gaussian kernels to different spatial and temporal scale parameters, while each convolution needs XYT multiplications of the size of the kernel used. This leads to the complexity of at least $O(XYT)$ which is the same as the complexity of our feature extraction method.

7. Conclusion

In this paper, we have studied the statistical properties of the 3D wavelet transform of videos and introduced a new approach to human action recognition and video activity level analysis based on these properties. Using the kurtosis values of marginal histograms, we have shown the non-Gaussian properties of these distributions. Also, the marginal histograms have been estimated by the GGD distributions. The study of joint statistics shows that the coefficients, conditioned on their parent, neighbors and cousins, are uncorrelated while dependent. The vertical cross sections of joint statistics indicate that the conditioned distributions are non-Gaussian as well. We have computed the MI, as a quantitative estimate of dependency, and have shown that the dependency increases when the activity in the video decreases. Moreover, the kurtosis curves have been proposed and grouped into four sets based on the degrees of activity in the video. Results show that kurtosis curves can reliably be used as indications of the activity level in the video. Finally, the joint and marginal statistical features of the 3D wavelet transform have been utilized for hierarchical SVM classification to determine human actions with 93.4% accuracy, outperforming the existing methods.

8. References

1. Turaga P., Chellappa R., Subrahmanian V.S., Udrea O. (2008) Machine Recognition of Human Activities: A Survey. IEEE Trans. on Circuits and Systems for Video Technology 18(11): 1473-1488. doi: 10.1109/TCSVT.2008.2005594

2. Sarkar S., Phillips P.J., Liu Z., Vega I.R., Grother P., Bowyer K.W. (2005) The humanID gait challenge problem: data sets, performance, and analysis. IEEE Trans. on Pattern Analysis and Machine Intelligence 27(2):162-177. doi: 10.1109/TPAMI.2005.39
3. Poppe R. (2010) A survey on vision based human action recognition. Image and Vision computing, Elsevier 28(6):976-990. doi:10.1016/j.imavis.2009.11.014
4. Moeslund T.B., Hilton A., Kruger V. (2006) A survey of advances in vision based human motion capture and analysis. Computer Vision and Image Understanding 104(2-3):90-126. 10.1016/j.cviu.2006.08.002
5. Bobick A.F., Davis J.W. (2001) The recognition of human movement using temporal templates. IEEE Trans. on Pattern Analysis and Machine Intelligence 23(3):257-267, 2001. doi: 10.1109/34.910878
6. Fathi A., Mori G. (2008) Action recognition by learning mid-level motion features. In: Proceeding of Int'l Conference on Computer Vision and Pattern, pp. 1–8
7. Ikizler N., Cinbis R.G., Duygulu P. (2008) Human action recognition with line and flow histograms. In: Proceedings of Int'l Conference on Pattern Recognition, pp. 1–4
8. Gorelick L., Blank M., Shechtman E., Irani M., Basri R. (2007) Actions as space–time shapes. IEEE Trans. on Pattern Analysis and Machine Intelligence 29(12):2247-2253. doi: 10.1109/TPAMI.2007.70711
9. Oikonomopoulos A., Patras I., Pantic M. (2006) Spatiotemporal salient points for visual recognition of human actions. IEEE Trans. on Systems Man and Cybernetics, Part B: Cybernetics 36(3):710-719. doi: 10.1109/TSMCB.2005.861864
10. Jhuang H., Serre T., Wolf L., Poggio T. (2007) A biologically inspired system for action recognition. In: Proceedings of Int'l Conference on Computer Vision, pp. 1-8
11. Laptev I., Marszałek M., Schmid C., Rozenfeld B. (2008) Learning realistic human actions from movies. In: Proceedings of Int'l Conference on Computer Vision and Pattern Recognition, pp. 1-8
12. Niebles J.C., Wang H., Fei L.F. (2008) Unsupervised learning of human action categories using spatial–temporal words. Int'l Journal of Computer Vision 79(3):299-318. doi: 10.1007/s11263-007-0122-4
13. Song Y., Goncalves L., Perona P. (2003) Unsupervised learning of human motion. IEEE Trans. on Pattern Analysis and Machine Intelligence 25(7):814-827. doi: 10.1109/TPAMI.2003.1206511
14. Kim T.K., Cipolla R. (2009) Canonical correlation analysis of video volume tensors for action categorization and detection. IEEE Trans. on Pattern Analysis and Machine Intelligence 31(8):1415-1428 doi: 10.1109/TPAMI.2008.167
15. Oikonomopoulos A., Pantic M., Patras I. (2009) Sparse B-spline polynomial descriptors for human activity recognition. Image and Vision Computing 27(12):1814-1825. doi: 10.1016/j.imavis.2009.05.010
16. Wong S.F., Kim T.K., Cipolla R. (2007) Learning motion categories using both semantic and structural information. In: Proceedings of Int'l Conference on Computer Vision and Pattern Recognition, pp. 1-8
17. Wong S.F., Cipolla R. (2007) Extracting spatiotemporal interest points using global information. In: Proceedings of Int'l Conference on Computer Vision, pp. 1-8
18. Schuldt C., Laptev I., Caputo B. (2004) Recognizing human actions: A local SVM approach. In: Proceedings of Int'l Conference on Patter Regocgnition, pp. 32-36
19. Laptev I., Lindeberg T. (2004) Local descriptors for spatio-temporal recognition. In: Proceedings of ECCV Workshop, Spatial Coherence for Visual Motion Analysis, pp. 91–103
20. Ngo C.W., Pong T.C., Zhang H.J. (2002) Motion-based video representation for scene change detection. Int'l Journal on Computer Vision 50(2):127–142. doi: 10.1023/A:1020341931699
21. Coudert F., Benois-Pineau J., Le Lann P.Y., Barba D. (1999) Binkey: A system for video content analysis on the fly, In: Proceedings of IEEE Int'l Conf. Multimedia Comput. Syst., 1:679–684
22. Nicolas H., Manaury A., Benois-Pineau J., Dupuy W., Barba D. (2004) Grouping video shots into scenes based on 1D mosaic descriptors. In: Proceedings of Int'l Conf. on Image Processing, ICIP, 1:637–640
23. Rajagopalan R., Orchard M.T. (2002) Synthesizing processed video by filtering temporal relationships. IEEE Trans. on Image Processing 11(1):26–36. doi: 10.1109/83.977880

24. Duan L.Y., Xu M., Tian Q., Xu C.S., Jesse S.J. (2005) A unified framework for semantic shot classification in sports video. IEEE Trans. on Multimedia 7(6):1066–1083. doi: 10.1109/TMM.2005.858395
25. Xu G., Ma Y.F., Zhang H.J., Yang S.Q. (2005) HMM-based framework for video semantic analysis. IEEE Trans. on Circuits Syst. Video Technol. 15(11):1422–1433. doi: 10.1109/TCSVT.2005.856903
26. Zhai Y., Shah M. (2006) Video scene segmentation using Markov chain Monte Carlo. IEEE Trans. on Multimedia 8(4):686–697. doi: 10.1109/TMM.2006.876299
27. Do M.N. (2001) Directional Multiresolution Image Representations. PhD thesis, Swiss Federal Institute of Technology
28. Cunha A.L., Do M.N., Vetterli M. (2007) A Stochastic Model for Video and its Information Rates. In: Proceedings of the 2007 Data Compression Conference, pp. 3-12
29. Lawrence Raniner R.R., Biing-Hwang J. (1993) Fundamentals of Speech Processing. Prentice-Hall International
30. Chen W., Zhang Y.J. (2008) Parametric model for video content analysis. Elsevier B.V., Pattern Recognition Letters. 29:181–191. doi: 10.1016/j.patrec.2007.09.020
31. Mo X., Wilson R. (2004) Video Modeling and Segmentation Using Gaussian Mixture Models. In: Proceedings of the 17th Int'l Conference on Pattern Recognition, ICPR 3:854-857
32. Greenspan H., Goldberger J., Mayer A. (2004) Probabilistic Space-Time Video Modeling via Piecewise GMM. IEEE Trans. on Pattern Analysis and Machine Intelligence 26(3):384-396. doi: 10.1109/TPAMI.2004.1262334
33. Li Z., Liu G. (2011) Video Scene Analysis in 3D Wavelet Transform Domain. Journal on Multimedia Tools and Applications. doi: 10.1007/s11042-010-0594-z
34. Mallat S. (1989) A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. on Patt. Recog. and Mach. Intell. 11(7):674– 693. doi: 10.1109/34.192463
35. Oh T.H., Besar R. (2003) JPEG2000 and JPEG: image quality measures of compressed medical images. In Proceedings of 4th National Conf. on Telecommunication Tech., pp. 31 – 35
36. Rapantzikos K., Avrithis Y.S., Kollias S.D. (2007) Spatiotemporal saliency for event detection and representation in the 3D wavelet domain: potential in human action recognition. CIVR 2007, pp. 294-301
37. Boashash B., (2003), Time-Frequency Signal Analysis and Processing: A Comprehensive Reference. Oxford, Elsevier Science.
38. DeVore R.A., Lucier B.J., Acta (1992) Wavelets. In: Proceedings of Numerica 92, A. Iserles, ed., Cambridge University Press, New York, pp. 1-56.
39. Meyer Y. (1989) Wavelets, In: Proceedings of Ed. J.M. Combes et al., Springer Verlag, Berlin, pp. 21
40. http://taco.poly.edu/WaveletSoftware/standard3D.html. Accessed 15 April 2011
41. Lian S., Sun J., Wang Z. (2004) A secure 3D-SPIHT codec. In: Proceedings of European Signal Processing Conference, pp. 813–816
42. Simoncelli E.P., Duccigrossi R.W. (1997) Embedded Wavelet Image Compression Based on a Joint Property Model. In: Proceedings of the IEEE Int'l Conf. On Image Processing 1:640-643. doi: 10.1109/ICIP.1997.647994
43. Moulin P., Liu J. (1999) Analysis of multiresolution image denoising schemes using generalized Gaussian and complexity priors. IEEE Trans. on Inform. Th. 45:909–919. doi: 10.1109/18.761332
44. Sharifi K., Leon-Garcia A., (1995) Estimation of shape parameter for generalized Gaussian distributions in subband decompositions of video. IEEE Trans. on Circuits Sys. Video Tech. 5:52–56. doi: 10.1109/76.350779
45. Wouwer G.V., Scheunders P., Dyck D.V. (1999) Statistical texture characterization from discrete wavelet representations. IEEE Trans. on Image Proc. 8(4):592–598. doi: 10.1109/83.753747
46. Po. D.D.-Y., Do M.N. (2003) Directional Multiscale Statistical Modeling of Images. Wavelets: Applications in Signal and Image Processing 5207:69-79. doi: 10.1117/12.506412
47. http://www.irisa.fr/vista/Equipe/People/Laptev/download.html. Accessed 15 September 2011
48. http://nsl.cs.sfu.ca/wiki/index.php/Video_Library_and_Tools. Accessed 15 September 2011
49. http://www.open-video.org. Accessed 15 April 2011
50. Po D.D.-Y., Do M.N. (2006) Directional Multiscale Modeling of Images using the Contourlet Transform. IEEE Trans. on Image Proc. 15(6):1610-1620. doi: 10.1109/TIP.2006.873450

51. Liu J., Moulin P. (2001) Information-theoretic analysis of interscale and intrascale dependencies between image wavelet coefficients. IEEE Trans. on Image Proc. 10(11):1647–1658. doi: 10.1109/83.967393
52. Cover T.M., Thomas J.A. (1991) Elements of Information Theory. Wiley Interscience. NewYork
53. Moddemeijer R. (1989) On estimation of entropy and mutual information of continuous distributions. Signal Proc. 16(3):233–246
54. Chang C.C., Lin C.J. (2001) LIBSVM : a library for support vector machines.
55. Recommendation ITU-R BT 500-6 (1994) Method for the subjective assessment of the quality of television pictures.
56. Omidyeganeh M., Ghaemmaghami S., Shirmohammadi S. (2010) Autoregressive Video Modeling through 2D Wavelet Statistics. In: Proceedings of the IEEE Int'l Conf. on Intelligent Information Hiding and Multimedia Signal Processing 1:272-275. doi: 10.1109/IIHMSP.2010.75
57. http://en.wikipedia.org/wiki/Wavelet. Accessed 14 September 2011
58. Do M.N., Vetterli M. (2000) Texture similarity measurement using Kullback-Leibler distance on wavelet subbands. In Proc. of IEEE Int'l Conf. on Image Processing 3:730-733. doi: 10.1109/ICIP.2000.899558
59. Kienzle V., Bakir G.H., Franz M.O., Schölkopf B. (2004) Efficient approximations for support vector machines in object detection. Pattern Recognition, Lecture Notes in Computer Science 3175:54-61. doi:10.1007/978-3-540-28649-3_7
60. Lu F., Yang X., Lin W., Zhang R., Yu S. (2011) Image classification with multiple feature channels Optical Engineering. 50(05). doi:10.1117/1.3582852
61. ITU-R Recommendation BT.500-11 (2002) Methodology for the subjective assessment of the quality of television pictures.
62. Simoncelli E. P., Portilla J. (1998) Texture characterization via joint statistics of wavelet coefficient magnitudes. In Proc. of IEEE Int'l Conf. on Image Processing 2:62-66. doi: 10.1109/ICIP.1998.723417.
63. Laptev I., Lindeberg T. (2004) Velocity adaptation of space-time interest points. In Proc. Of In'l Conf. on Pattern Recognition 1:52- 56. doi: 10.1109/ICPR.2004.1334003
64. Sun X., Chen M., Hauptmann A. (2009) Action recognition via local descriptors and holistic features. In Proc. Of IEEE Int'l Conf. on Computer Vision and Pattern Recognition Workshops 58-65. doi: 10.1109/CVPRW.2009.5204255