Spatiotemporal saliency for event detection and representation in the 3D Wavelet Domain: Potential in human action recognition

Konstantinos Rapantzikos, Yannis Avrithis and Stefanos Kollias
School of Electrical & Computer Engineering
National Technical University of Athens
Iroon Polytexneiou 9, 15780 Zografou, Greece
Tel: +30-210-7724351
e-mail: {rap,iavr}@image.ntua.gr, stefanos@cs.ntua.gr

ABSTRACT
Event detection and recognition is still one of the most active fields in computer vision, since the complexity of the dynamic events and the need for computational efficient solutions pose several difficulties. This paper addresses detection and representation of spatiotemporal salient regions using the 3D Discrete Wavelet Transform (DWT). We propose a framework to measure saliency based on the orientation selective bands of the 3D DWT and represent events using simple features of salient regions. We apply this method to human action recognition, test it on a large public video database consisting of six human actions and compare the results against an established method in the literature. Qualitative and quantitative evaluation indicates the potential of the proposed method to localize and represent human actions.

Categories and Subject Descriptors
I.4.8 [**Scene Analysis**]

I.4.7 [**Feature Measurement**] – *Feature representation*

General Terms
Algorithms

Keywords
Spatiotemporal saliency, action recognition, 3D wavelet transform

1. INTRODUCTION
As collections of video data grow automatic systems for efficiently querying this data about abstract or specific events have become a significant need. Dynamic event detection and representation is not straightforward even for videos with simple temporal structure. Most approaches in the field are either model-based [1][3][4], attempting to estimate a set of model parameters from the video data, or appearance- based [2][5][6][7], that perform inference directly on the observed data. In order to reach a semantic level of event description there is no need to process all available visual information. Specific parts of the scene are usually representative enough so that limiting further analysis to them reduces complexity and enhances understanding. This is the main idea of most approaches that extract interest/salient points or regions. Those areas are usually located around corners, edges or highly textured regions. Such Regions-Of-Interest (ROI) may either be defined directly, after fusing/combining different features extracted from the image or indirectly, after extracting interest points and grouping them together.

The majority of salient point detectors is based on forming matrices that describe the gradient distribution in a local neighborhood of a point. The eigenvalues of these matrices represent the main neighborhood directions and are enough for measuring saliency so that if a significant change occurs, then this is one of the candidate points of interest. Lindeberg [21] proposed a Hessian-affine detector that shares a similar idea, since the second derivatives involved in the Hessian matrix give strong response on blobs and ridges. This method becomes scale invariant after selecting the characteristic scale of a local structure, for which a given function (e.g. a Laplacian) attains an extremum over scales [20]. Given the set of initial points extracted at their characteristic scales, an iterative estimation of elliptical affine region is applied in order to obtain the desired ROI. Affine region detectors have reached some maturity level in the computer vision literature. The detection of regions covariant with a class of affine transformations has been used in many applications including large scale image retrieval [8][9], object retrieval in video [10][11], texture recognition [4], object categorization [12][13] and symmetry detection [14]. Six methods for detecting such regions in images are described and evaluated in the recent work of Mikolajczyk *et al.* [15]. Overall, first the interest points are detected in scale-space with one of the methods and then an elliptical region for each of them is defined. Finally, proper selection and grouping provides the desired ROI.

While a large amount of work has been done on representing spatial information, far less work has been done in exploiting the spatiotemporal video structure to detect video activities. Bobick *et al.* [2] construct Motion-History-Images (MHI) for representing human actions and use moment invariants to represent them. Although efficient, this method requires that the object performing the action is well segmented from the background.

CIVR'07, July 9-11, 2007, Amsterdam, The Netherlands.

Most of the methods that fall into this category are extensions of the ones discussed before and exploit the temporal video structure using small spatiotemporal neighborhoods for detecting/selecting points of interest. The majority of these techniques is applied and tested to human action recognition. Laptev *et al.* [16] [17] and Schuldt *et al.* [18] build on the idea of Harris & Forstner interest point operators and propose a method to detect points that correspond roughly to points in space-time where motion abruptly changes direction. They adapt those points to velocity and scale and show how they correspond to interesting events by applying them to video interpretation and human action recognition [18]. Ke *et al.* extract volumetric features from spatiotemporal neighborhoods and construct a real-time event detector for complex actions of interest with quite promising results [7]. Boiman *et al.* [19] and Zelnik-Manor *et al.* [5] have used overlapping volumetric neighborhoods for analyzing dynamic actions, detecting salient events and detecting/recognizing human activity. In our earlier work in Rapantzikos *et al.* [23][24][25], we have proposed a volumetric framework for computing saliency based on intensity, color and orientation with applications to video ROI detection and video classification. This framework was based on the work of Itti *et al.* for visual attention on static images [26][27], which has proven its efficiency in several computer vision applications. Generally, all methods show the positive effect of using spatiotemporal information in video event analysis applications.

Although the notion of saliency remains the same, in this paper, we present a new framework, when compared to our previous work, for spatiotemporal salient regions detection and representation and test its efficiency on a human action recognition application. In this framework, a video sequence is represented as a solid in the three-dimensional Euclidean space, with time being the third dimension. We apply the multiscale 3D wavelet transform to decompose the volume into orientation sensitive subbands and use the resulting coefficients to compute saliency, detect regions of importance and extract representative features. The assumption of our method is similar to the one of most researchers in the field. We assume that the neighborhoods around points where significant motion, i.e. 3D orientation, change occurs are most important for interpreting dynamic actions. We also incorporate simple geometric constraints for boosting performance. The main focus of this paper is on exploring the ability of the 3D wavelet transform to locate and represent dynamic events with the constraint of keeping the computational complexity low, while performing efficiently.

The proposed method is compared against the one proposed by Laptev *et al.* in [16][17] for a human action retrieval application. We use a public and well structured database of video clips showing people performing several different actions for presenting statistics and evaluating the techniques.

The paper is organized as follows. Section 2 provides an introduction to the 3D wavelet transform in order to get a better insight of its orientation selectivity, while section 3 describes the methodology for computing saliency in the wavelet domain. In section 4 we provide extended statistics on the human action database by comparing the proposed method with a state-of-the art one and in section 5 we draw conclusions and discuss future work.

2. WAVELET-BASED SPATIOTEMPORAL VIDEO REPRESENTATION

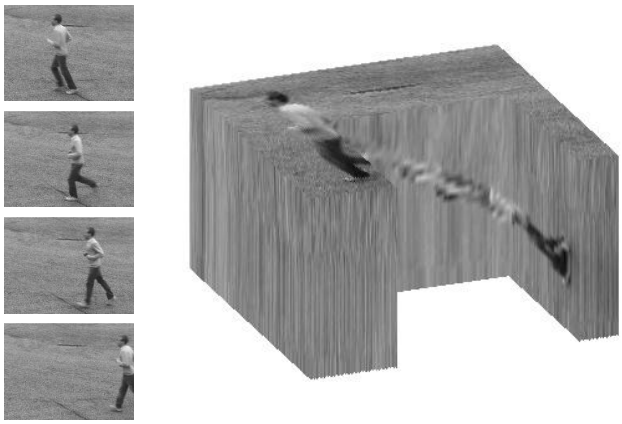

Fig. 1 Volumetric representation of a jogging sequence (the central part is carved out)

2.1 Spatiotemporal Video Representation

Given an arbitrary input sequence, we form a volume in a discrete 3D space, which is a set of grid points in 3D Euclidean space defined by their Cartesian coordinates (*x, y, t*). Specifically, the spatial dimensions of width and height are the *x*- and *y*- axes of a frame, while the temporal one is derived by layering the frames sequentially in time (*x*-*y*-*t* space). The minor element of this volumetric representation is called voxel and is defined as the unit cubic volume centered at the integral grid point. Such a volumetric representation provides richer information about the video structure along a large temporal scale than the individual 2D frames.

A moving object in such a volume is perceived as occupying a 3D region in space-time volume. Fig. 1 shows the volumetric representation of a jogging sequence with a central part of the volume being carved out.

2.2 3D Discrete Wavelet Transform

The 3D DWT is a separable transform that can be constructed by three separate 1D wavelet transforms. The signal at each of the three dimensions is convolved with a low-pass *L* and a high-pass *H* filter and downsampled by two. If we define $F(x,y,t)$ to be the volume corresponding to an input sequence, then *F* is initially filtered along the *x*-dimension, resulting in a low-pass volume $L(x,y,t)$ and a high-pass one $H(x,y,t)$. These volumes are subsampled by two and filtered along the *y*-dimension, resulting in four subvolumes, namely *LL*, *LH*, *HL*, *HH*. Then the subvolumes are downsampled once again and filtered along the *t*-dimension, resulting in eight subvolumes, namely *LLL*, *LLH*, *LHL*, *LHH*, *HLL*, *HLH*, *HHL* and *HHH*. Hence, the result of the decomposition is a multiscale structure of an approximation subband and a series of detail subbands w_i^l, where *l* denotes the number of scales and $i \in \{LLH, LHL, LHH, HLL, HLH, HHL, HHH\}$. Fig. 2 illustrates the 3D decomposition of a volume.

The subbands at each level of the decomposition have certain properties related to spatiotemporal orientations in the input

volume due to the frequency ranges they contain. For example, the *LLL* subvolume corresponds to the slowly moving average signal (approximation band), *LLH* emphasizes the quickly changing average signal, *LHL* the slowly changing horizontal, *HLL* the slowly changing vertical, *HHH* the quickly changing diagonal features etc. Fig. 3 depicts a slice of the wavelet decomposition of a running sequence for each of the bands. Light values correspond to strong coefficients.

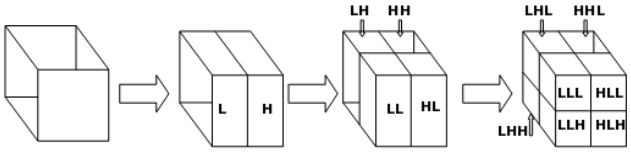

Fig. 2 intermediate steps for computing the 3D Discrete Wavelet Transform

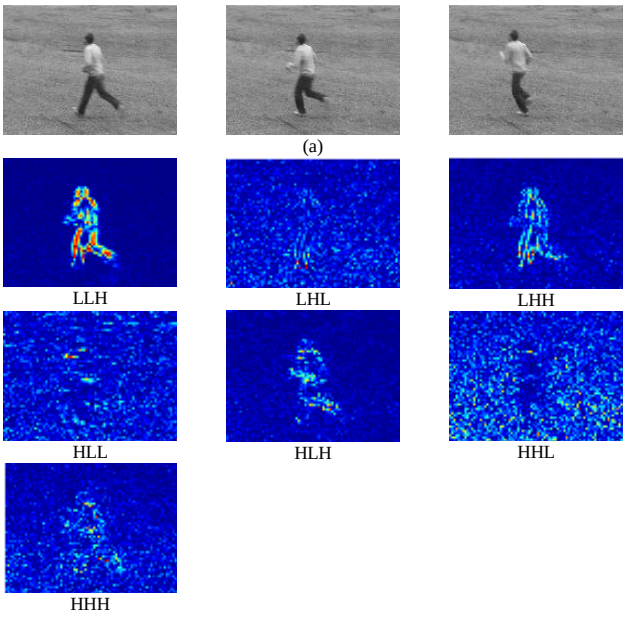

Fig. 3 (a) three frames of a "running" sequence; (b) the 7 detail bands of 3D-DWT for the middle frame of (a). The images are better viewed in color, where red corresponds to high values and while blue to low ones.

3. SALIENT REGION DETECTION AND REPRESENTATION

3.1 Saliency estimation

As discussed before, the wavelet transform of a clip gives information about spatiotemporal frequencies of the signal, while localizing them in space, time and scale. These frequencies correspond to different spatiotemporal orientations and therefore obtaining such a high joint resolution in space/time and spatiotemporal frequency is critical for detecting and analyzing dynamic events. We expect that spatiotemporal regions of interest will pop out in the transformed domain due to the high energy concentration in one or more bands. Exhibiting high energy concentration in more than one band means that this region becomes more salient, since it corresponds to a region where a significant motion change occurs.

Following this rationale, we assume that dynamic video content can be characterized by measuring the distribution of spatiotemporal energy across the 3D wavelet subbands and we define a voxel or a spatiotemporal neighborhood as salient, if it exhibits high inter- and intra- band energy concentration across scales. We realize this assumption by computing the energy of each voxel in a small neighborhood at each band and fusing it with the same neighborhoods across bands. In the wavelet domain, this kind of fusion becomes quite simple, since strong signal coefficients are already clearly separated from the rest and noise is spread all over the bands. Experimentation with various simple fusion techniques prove that addition of the computed energies across bands is enough to make salient regions pop out and measure the amount of saliency for each coefficient, while keeping computational complexity low. This measure may be also considered as a measure of spatiotemporal texture over time. We compute local energy for each subband by

$$E^l_i(x,y,t) = \frac{1}{|N|} \sum_{\{x,y,t\}\in N} |N(w_i(x,y,t))|^2 \quad (1)$$

where N is a local neighborhood defined around each wavelet coefficient, l the number of scales and i the subband index as defined in section 2.2. The resulting energies for six human actions and for $3\times3\times3$ neighborhood N around each coefficient are shown in Fig. 4, 5. Notice the well defined regions around the areas of interest that are important for interpreting each of the actions (e.g. hands, knees, legs, etc.).

Finally the saliency measure is computed by

$$S(x,y,t) = \sum_i E^l_i(x,y,t), \quad l=1 \quad (2)$$

Notice that in this work we only used the 1st scale of the wavelet transform, which is the finer one. More refined results in case of complex videos may be obtained by exploiting the other scales too, but it was not necessary for the data we processed. A further insight in the nature of the fused bands and the computed saliency is given in the next section.

3.2 Representation of salient regions

As mentioned in the introduction, the motivation of this work was to propose a computationally efficient but successful technique to represent spatiotemporal actions. Nevertheless, we started to think towards this direction after visually exploring the strength of the 3D DWT to represent complex actions in its subbands. After decomposing a small number of different human actions videos, it became evident that local activation of groups of specific subbands seemed to correspond to specific actions.

In Fig. 4, 5 one frame for each of six human action sequences are depicted along with the corresponding slice of the local energy computed for each of the wavelet subbands. These sequences are from a public available dataset [22] and correspond to *boxing*, *hand-clapping*, *hand-waving*, *jogging*, *running* and *walking*. More details about the data will be given in the experimental section. Just a glance at the figures reveals the differences in subband activity after decomposing each action sequence and computing the energies (Eq. 1). For example the LLH band in the boxing, hand-clapping and hand-waving sequences is different both in terms of magnitude and geometrical configuration of high activity areas. The same holds for the LHH band. Similar observations can be made from the slices in Fig. 5 that contain the actions with higher motion activity.

Since one of our goals was to explore the potential of the wavelet coefficients in representing complex actions, we experimented with different combinations of the bands for the saliency computation in Eq. 2. Fig. 6 shows results on a handwaving sequence after combining two sets of bands, namely the $b1 = \{LLH, LHH, HLH, HHH\}$ and $b2 = \{LHL, HLL, HHL\}$. Set *b1* is more related to fast changing regions, which are usually related to the foreground, while b2 is related to slowly changing regions related to the background. This difference is quite obvious for actions of low motion activity like boxing, hand-clapping and hand-waving as can be seen in Fig. 6. Fig. 6b depicts the result after fusing only the subbands in *b1* and Fig. 6c after fusing the

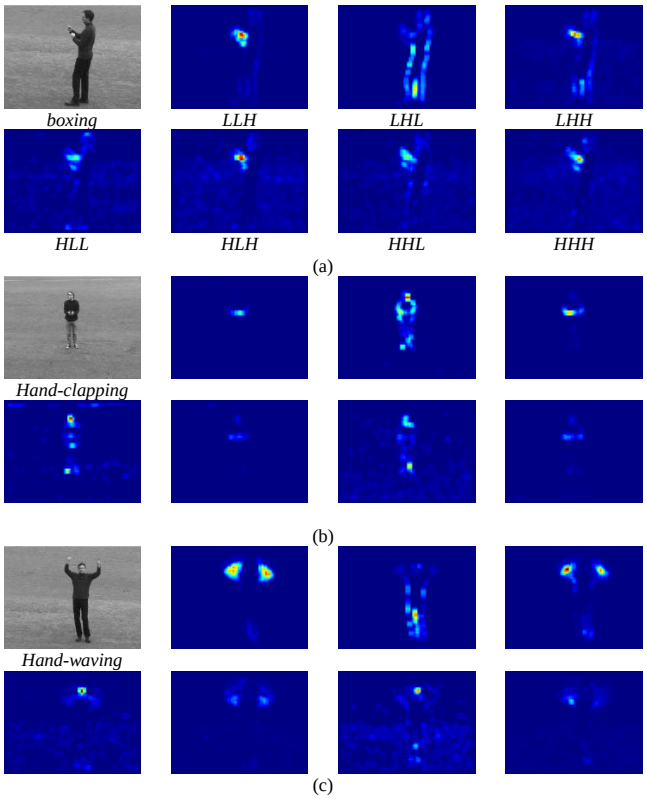

Fig. 4 the 7 wavelet subbands corresponding to a frame of a (a) boxing, (b) hand-clapping and (c) hand-waving sequence. (bands' labels are shown in (a))

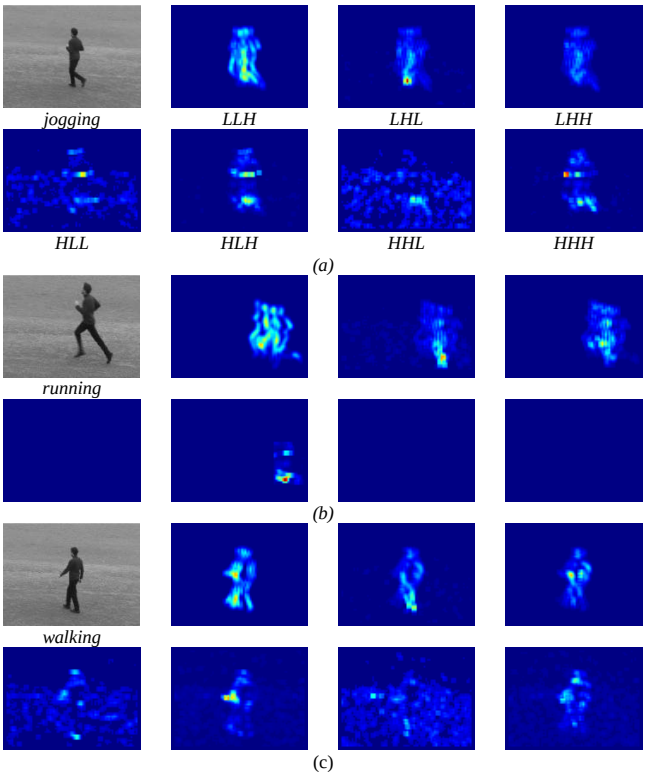

Fig. 5 the 7 wavelet subbands corresponding to a frame of a (a) jogging, (b) running and (c) walking sequence. (bands' labels are shown in (a))

subbands in *b2*. These actions do not include significant torso or body motion. Nevertheless, the difference is lower when looking at the results for the rest of the actions. It seems reasonable to weight the contribution of these two sets according to the activity in the scene in order to obtain more representative regions, but in this paper we select to use all of the bands interchangeably, since we are rather focusing on the potential of the spatiotemporal wavelet domain to detect ROIs rather than optimal performance.

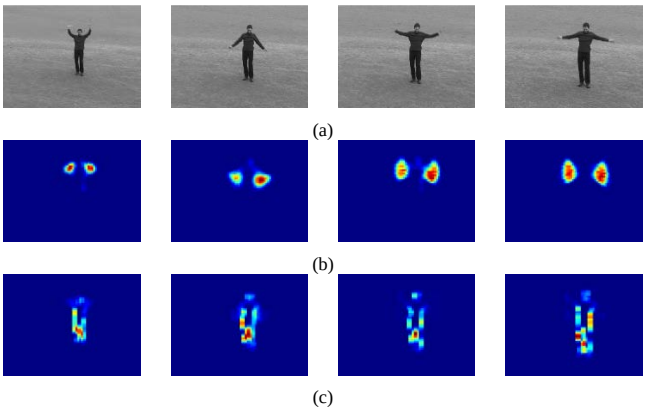

Fig. 6 (a) neighboring frames from a hand-waving sequence under camera zoom in/out; (b) saliency using bands in b1; (c) saliency using bands in b2; (see text)

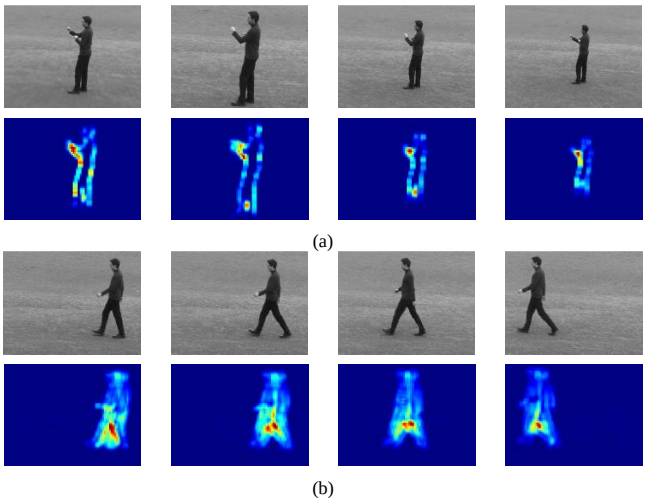

Fig. 7 (a) neighboring frames from a boxing sequence under camera zoom in/out and the corresponding saliency using all bands; (b) the same for a walking sequence.

The geometry of an action, as supported by researchers in the field, plays an important role in recognition. As mentioned in the introduction the shape descriptions of the extracted MHI silhouettes of Bobick *et al.* were successful, but required exact segmentation [2]. Boiman *et al.* on the other hand have used loose geometrical constraints to enhance the known event detection ability of their method [19]. In an attempt to incorporate such constraints in our framework, while keeping computational resources low, we run a *k*-means algorithm to detect a number of clusters on the saliency volumes and select the *p* most populated ones. In this way we obtain *p* centroids that correspond to the most salient areas of the mask. Visual examination of the results proves that this clustering leads to *p* points on the most important parts of the input, which correspond to e.g. the two hands in a hand-waving sequence, to the legs area for a walking one or the neck and legs area of a running sequence.

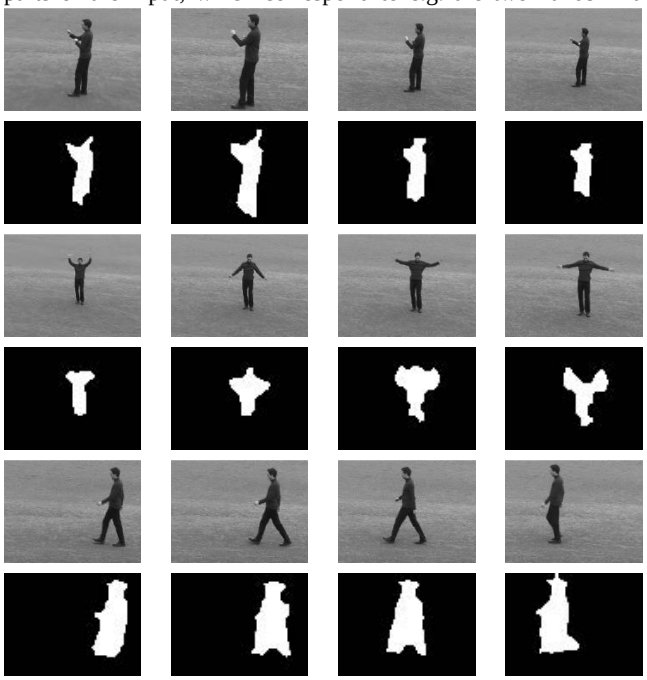

Fig. 8 Binary masks used to enhance the computational efficiency of the clustering step (see text).

In order to reduce computational complexity of the clustering algorithm by limiting the number of observations, we threshold the saliency volumes using an automatic method [28] and obtain binary masks. It is worth mentioning that this step is not crucial for the method, since the salient areas are already clearly defined and there is no possibility to threshold out areas of interest. Examples for these masks are given in Fig. 8. Notice that the obtained masks can be interpreted as motion history masks, similar in a way to the MHI discussed before, since they contain the accumulated motion of the neighboring frames in the neighborhood *N* used to compute local energies in Eq. 1. Connected *K*-means is then applied having as variables the saliency value of each voxel in the mask and the corresponding *x*, *y*, *z* coordinates. The centroids are normalized with respect to the centroid of the mask and the histograms are normalized with respect to standard deviation and mean value.

The final feature vector *F* is obtained by combining all computed features from the binary mask as

$$F = \{H_p^l(E_i),\ C_p^l\} \tag{3}$$

where H_n denotes the *n*-bin histogram of each energy band, C_p the selected *p* points after the k-means clustering for *l* levels of the wavelet transform. Histograms are computed only for the voxels in the binary masks and the first bin that corresponds to the least salient voxels is discarded.

4. EXPERIMENTS

4.1 Experimental setup and methodology

For evaluating the proposed framework, we select human action recognition, set up an action retrieval application and compare the results against the method proposed by Laptev *et al.*, which is available on-line [16]. We used a public database to evaluate both methods [22]. This database consists of six types of human actions (walking, jogging, running, boxing, hand-waving, hand-clapping) performed by 25 subjects in four different scenarios: outdoors s1, outdoors with scale variation s2, outdoors with different clothes s3 and indoors s4. All sequences were recorded with a static camera at a 25fps rate and have a size of 160x120. Annotation at the frame level is also available indicating when an action starts and when it stops. In this way, each video is split into four sub-sequences of almost equal length. For our experiments, we use the first 32 frames of each sub-sequence as a representative clip of the action.

Laptev *et al.* [16] and Schuldt *et al.* [18] detect local space-time interest points, adapt them to position, size and velocity of the moving patterns and extract spatiotemporal jets of order four around each point. Their implementation returns all points-of-interest without any selection, but in [16] they propose to run a k-means clustering and select four candidate interest points from the four most populated clusters. We implemented this step in order to be fair and avoid lengthy feature vectors. Hence, the length of the feature vector for this method is 136, since 4 points are selected and 34 local spatiotemporal Harris jets are computed for each of them. The feature vector of our method is lengthier, since we use *n=7*, *p=3* and *l=2* in (3). This gives rise to a feature vector of length 224.

For performance evaluation we applied the leave-one-out approach to a similarity retrieval application, taking each action clip one after the other, removing its contribution from the database, finding its action label and comparing the result to its actual label. For similarity we used the Euclidean distance, which for the representation of two sequences k_1 and k_2 is defined by

$$d_E(F_{k_1}, F_{k_2}) = \sqrt{\sum_{j=1}^{L} (F_{k_1}(j) - F_{k_2}(j))^2} \quad (4)$$

where L denotes the length of the feature vector.

4.2 Results

In this section we present statistics on retrieving actions from the available database. Tables 1, 2 show results on the s1 scenario for the proposed method with and without geometric constraints, with the last method being more successful in identifying the actions. Notice the improvement in differentiating hand-waving from boxing and hand-clapping/walking from boxing sequences. The first is mainly due to the different geometry of the raising arms, while the second mainly due to the closeness of the salient centroids (upper legs part) to the mask centroid. Table 3 shows the retrieval results when using the Laptev *et al.*'s method, which are lower than the corresponding results in tables 1 and 2. Notice the high improvement achieved by our technique in detecting jogging sequences and differentiating them from the running ones when compared to the other method. These two actions are the most confusing ones, as Schuldt *et al.* mention in their action classification application [18].

Table 1 Proposed method on scenario s1

	Box	Hclp	Hwav	Jog	Run	Walk
Box	**64**	10	17	0	0	8
Hclp	18	**59**	20	0	0	3
Hwav	15	12	**72**	0	0	1
Jog	0	0	1	**74**	10	15
Run	0	0	0	22	**75**	3
Walk	12	6	3	14	3	**62**
prec	0.587	0.678	0.637	0.673	0.852	0.674
rec	0.646	0.590	0.720	0.740	0.750	0.620

Table 2 Proposed method with geometric constraints on scenario s1

	Box	Hclp	Hwav	Jog	Run	Walk
Box	**73**	14	11	0	0	1
Hclp	15	**70**	12	0	0	3
Hwav	5	6	**85**	0	0	4
Jog	0	1	0	**74**	15	10
Run	0	0	0	22	**75**	3
Walk	5	5	3	14	1	**72**
prec	0.745	0.729	0.766	0.673	0.824	0.774
rec	0.737	0.700	0.850	0.740	0.750	0.720

Table 3 Laptev *et al.* method adapted, Confusion matrix, scenario s1

	Box	Hclp	Hwav	Jog	Run	Walk
Box	**45**	15	7	9	15	8
Hclp	15	**44**	11	3	13	14
Hwav	5	8	**57**	4	14	12
Jog	1	1	9	**32**	33	24
Run	1	6	1	11	**75**	6
Walk	1	7	8	21	19	**44**
prec	0.662	0.543	0.613	0.400	0.444	0.407
rec	0.455	0.440	0.570	0.320	0.750	0.440

For the sake of completeness, we also include results of our method for all sequences involved in scenarios s1, s2 and s3. As derived from the statistics in Table 4, it performs adequately, without any specific adaptations for facing e.g. the scale change in s2, which is the most difficult scenario as Shuldt *et al.* mention in [18].

Computational efficiency was also one of our goals. Indicatively, for a Pentium IV, 2.4GHz and 512 MB RAM, the average processing time of the proposed method is 19s, while for Laptev *et al.*'s method is 319s. The maximum number of iterations to achieve adaptation convergence for the Laptev *et al.*'s method is set to 20 as in the public code. Both implementations are in MATLAB. Processing time for our method depends on the size of the obtained binary mask, while for the Laptev *et al.*'s method the convergence or non-convergence of scale-velocity adaptation is quite critical. Both methods are more computationally demanding when dealing with the more dynamic actions of the database, namely jogging, running and walking. In order to be fair, we should mention that the performance of Laptev *et al.*'s method depends on the number of points that adapt successfully to scale and velocity. The adaptation, in turn, depends on the defined number of max iterations. Nevertheless, setting a high value for this number increases dramatically the computational time.

Table 4 Proposed method, Confusion matrix, with geometric constraint, scenarios s1, s2, s3

	Box	Hclp	Hwav	Jog	Run	Walk
Box	**194**	47	32	7	3	16
Hclp	70	**154**	36	11	8	17
Hwav	37	39	**197**	7	1	17
Jog	7	9	3	**168**	70	43
Run	5	7	7	80	**180**	21
Walk	22	23	27	46	21	**161**
prec	0.579	0.552	0.652	0.527	0.636	0.585
rec	0.649	0.520	0.661	0.560	0.600	0.537

5. CONCLUSIONS AND FUTURE RESEARCH

In this paper we propose a framework for spatiotemporal saliency computation in the 3D wavelet domain, in order to represent human motion. We use a set of wavelet-based and geometric

features that correspond to intra- and inter- activity peaks in the orientation sensitive wavelet subbands and give a deep insight of the abilities of the 3D DWT to locate dynamic events in the input sequence. The efficiency of the method in recognizing human actions is illustrated by comparison against a well established technique on a public video dataset consisting of six actions.

This paper should be regarded as an attempt to illustrate the strengths in terms of simplicity and computational efficiency of the wavelet transforms to represent dynamic video events. In the future, we wish to elaborate on saliency computation by automatic weighted fusion of bands in order to deal with more complex sequences. We will also consider using the more orientation selective 3D Dual-Tree Wavelet Transform [29] for the same goal. Additionally we wish to increase the discriminative power of the proposed method by investigating the optimality in selecting a number of spatiotemporal points from the saliency volume and extracting features more robust to scale. Specifically, for human action recognition, we will focus on improving the statistics shown in Table 4. In general we will examine ways to use the method for general video event detection and representation in a computationally efficient way.

6. ACKNOWLEDGMENTS

This research work was supported (in part) by MUSCLE, FP6-507752 and MESH, FP6-027685

7. REFERENCES

[1] Ramanan, D., Forsyth, D. A. Using temporal coherence to build models of animals. In Proc. ICCV, vol. 1, pp. 338-345, Oct 2003.

[2] Bobick, A.F., Davis, J.W. The recognition of human movement using temporal templates. IEEE Trans. PAMI, vol. 23, pp. 257-267, 2001.

[3] Schmid, C. Constructing models for content-based image retrieval. In roc. CVPR, vol. 2, pp. 39-45, 2001

[4] Lazebnik, S., Schmid, C., Ponce, J. Affine-invariant local descriptors and neighborhood statistics for texture recognition. in ICCV, vol. 1 , 649- 655, 2003.

[5] Zelnik-Manor, L., Irani, M. Statistical Analysis of Dynamic Actions. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), vol. 28, no. 9, pp. 1530--1535, Sep 2006.

[6] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, Actions as Space-Time Shapes. IEEE International Conference on Computer Vision (ICCV), Beijing, October 2005.

[7] Y. Ke, R. Sukthankar, M. Hebert. Efficient Visual Event Detection using Volumetric Features. International Conference on Computer Vision, vol. 1, pp. 166–173, Oct 2005.

[8] C. Schmid, R. Mohr. Local gray value invariants for image retrieval. In. IEEE Trans. On Pattern Analysis and Mach. Intelligence, vol. 19, no. 5, pp. 530-535, 1997.

[9] T. Tuytelaars,L. VanGool, L. D'haene, R. Koch. Matching of affinely invariant regions for visual servoing. In Proc. Int. Conf. Robotics and Automation, ICRA'99

[10] J. Sivic, A. Zisserman.Video google: A text retrieval approach to object matching in videos. In Proc. International Conf. on Computer Vision (ICCV03), vol. 2, pp. 1470, 2003.

[11] J. Sivic, F. Schaffalitzky, A. Zisserman. Object level grouping for video shots. In Proc. of the 8th European Conf. on Computer Vis., pp. 724-734, 2004.

[12] G. Csurka, C. Dance, C. Bray, L. Fan..Visual categorization with bags of keypoints. In Proc. Statistical Learning in Computer Vision, 2004

[13] A. Opelt, M. Fussenegger, A. Pinz, P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In Proc. of European Conference on Computer Vision, Prague, Czech Republic, pp. 71–84, 2004

[14] A. Turina, T. Tuytelaars, L. VanGool. Efficient Grouping under perspective skew. In Proc. IEEE Conf. on Computer Vis. and Patt. Recogn. (CVPR01), 2001.

[15]] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, L. Van Gool. A comparison of affine region detectors. In Intern. Journal of Computer Vision, vol. 65, no. 1, pp. 43-72, 2006

[16] I. Laptev and T. Lindeberg. Space-Time Interest Points. In Proc. ICCV'03, Nice, France, pp. 432-443, 2003 (matlab implementation in http://www.nada.kth.se/~laptev/code.html)

[17] I. Laptev, T. Lindeberg , "Local Descriptors for Spatio-Temporal Recognition", ECCV Workshop "Spatial Coherence for Visual Motion Analysis", May 2004

[18] C. Schuldt, I. Laptev, B. Caputo , "Recognizing Human Actions: A Local SVM Approach", in Proc. ICPR'04, 2004

[19] O. Boiman and M. Irani, "Detecting irregularities in images and in video", IEEE International Conference on Computer Vision (ICCV), Beijing, Oct 2005

[20] T. Lindeberg, J. Garding. Shape-adapted smoothing in estimation of 3-D shape cues from affine deformations of local 2-D brightness structure. In Image and Vision Computing, vol. 15, no.6, pp. 415–434, 1997

[21] T. Lindeberg. Feature detection with automatic scale selection. In International Journal of Computer Vision, vol. 30, no.2, pp. 79-116, 1998

[22] Human action dataset, http://www.nada.kth.se/cvap/actions/

[23] K. Rapantzikos, N. Tsapatsoulis, Y. Avrithis, "Spatiotemporal Visual Attention Architecture for Video Analysis Proc. of IEEE International Workshop On Multimedia Signal Processing (MMSP'04), Sienna, 2004

[24] K. Rapantzikos, N. Tsapatsoulis, Y. Avrithis, S. Kollias, "A bottom-up spatiotemporal visual attention model for video analysis", IEE Vision, Image and Signal Processing, accepted for publication

[25] K. Rapantzikos, Y. Avrithis, "An enhanced spatiotemporal visual attention model for sports video analysis", International Workshop on Content-based Multimedia indexing (CBMI'05), Riga, Latvia, Jun 2005

[26] L. Itti, C. Koch , "A saliency-based search mechanism for overt and covert shifts of visual attention", Vision Research, vol. 40, pp. 1489-1506, 2000

[27] L. Itti, C. Koch, E. Niebur, "A model of saliency-based visual attention for rapid scene analysis", IEEE Trans. on

Patt. Analysis and Mach. Intell., vol. 20, no. 11, pp. 1254-1259, 1998

[28] Otsu N., "A thresholding selection method from gray-scale histogram," IEEE Trans. on System, Man, and Cybernetics, vol. 9, pp. 62-66, 1979.

[29] N.G. Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals", *Applied Computational Harmonic Anal*., vol. 10, no. 3, pp. 234-253, May 2001