Recognition of Activities of Daily Living

Konstantinos Avgerinakis
Information Technologies Institute
Centre for Research and Technology, Hellas
Thessaloniki, Greece
Email: koafgeri@iti.gr

Alexia Briassouli
Information Technologies Institute
Centre for Research Technology, Hellas
Thessaloniki, Greece
Email: abria@iti.gr

Ioannis Kompatsiaris
Information Technologies Institute
Centre for Research Technology, Hellas
Thessaloniki, Greece
Email: ikom@iti.gr

***Abstract*—This paper presents a new method for human action recognition which exploits advantages of both trajectory and space-time based approaches in order to identify action patterns in given sequences. Videos with both a static and moving camera can be tackled, where camera motion effects are overcome via motion compensation. Only pixels undergoing changing motion, found by extracting motion boundary-based activity areas, are processed in order to introduce robustness to camera motion and reduce computational complexity. In these regions, densely sampled grid points on multiple scales are tracked using a KLT tracker, leading to dense multi-scale trajectories, on which HOGHOF descriptors are estimated. The length of each trajectory is determined by detecting changes in the tracked points' motion or appearance using sequential change detection techniques, namely the CUSUM approach. A vocabulary is created for each video's features using Hierarchical K-means, and the resulting fast search trees are used to describe the actions in the videos. SVMs are used for classification, using a kernel based on the similarity scores between training and testing videos. Experiments are carried out with new and challenging datasets for which the proposed method is shown to lead to recognition results that are comparable to or better than existing state of the art methods.**

I. Introduction

The problem of human activity recognition is one of the central ones in computer vision, attracting significant research attention. Numerous methods have been developed for it, with space-time feature based approaches being among the most popular ones for video representation, which, in combination with Bag of Features methods, lead to State of the Art results. Early approaches were focused either on the estimation of trajectories for the description of the actions, or were based on the extraction of space-time volumes [1] such as silhouettes and motion history images [2]. However, they lacked robustness to commonly encountered videos with moving cameras, and changes in scale or orientation, for which local feature-based approaches have proven to be more suitable. Feature-based methods have been used for activity recognition because of their success in object recognition and numerous types of features have been developed and tested for human activity recognition [3]. Usually action recognition is based on sparse features in order to reduce the computational cost, but results in performance degradation as information can be lost due to the small number of features. Recent methods are focusing on the use of dense interest point tracking [4], as it provides more information about the scene and, consequently, more accurate recognition results.

II. Motivation, Contribution

Our approach has been motivated by the very good results presented in [3], where dense trajectories are used for action recognition by applying dense sampling with image pyramids, and a simple Optical Flow (OF)-based tracking algorithm for the construction of trajectories. In our work, dense multi-scale interest points are detected and tracked, but inspired by [5], we use a KLT tracker and combine it with a RANSAC estimator to eliminate the erroneous KLT predictions and improve the tracking process at no additional computational cost. In [3], Motion Boundary Histogram (MBH) descriptors are used to characterize the tracked grid points, as they are more robust to camera motion than Histograms of Optical Flow (HOFs). Their approach is based on the optical flow estimates of [6] which provide more accurate MBHs than HOFs. In our work, we use the optical flow of [7], as it provides a better localization of pixels undergoing a changing motion, which is necessary for the reduction of the method's computational cost and reduction of false alarms (see Sec. III). After extensive experimentation with these flow estimates, the combination of Histograms of Oriented Gradients (HOGs) with HOFs, referred to as HOGHOFs, are shown to lead to more accurate recognition results, and are therefore preferred over the MBHs. In order to use HOGHOFs without inaccuracies caused by camera motion, we introduce a preprocessing motion compensation step [8]. Unlike [3] who process all frame pixels, in each frame we compute a region of pixels whose motion is changing based on motion boundary statistics, and apply dense sampling to it. This minimizes the amount of data that is sampled and consequently reduces the computational cost of the tracking procedure, while also eliminating the need to remove static pixels from the estimated trajectories later on in the tracking process.

A singificant innovation that we introduce is the creation of trajectories with a meaningful temporal extent. In contrast to existing work, e.g. [3], we do not simply use a predetermined trajectory length. Instead, we apply Cumulative Sum based sequential change detection (CUSUM) on the tracked points and find changes in their motion descriptors (HOFs) or appearance descriptors, namely the Histograms of Oriented Gradients (HOGs), automatically limiting the trajectory in a meaningful manner related to actual changes in activity motion or appearance. The interest points that are tracked

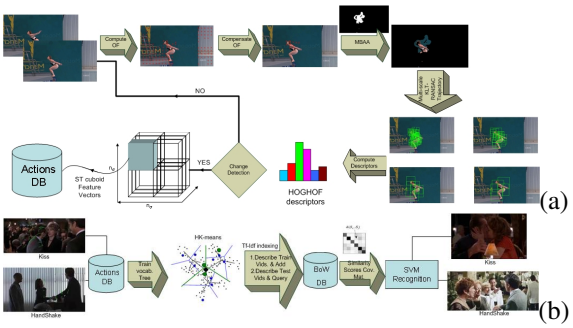

Fig. 1. Block diagrams of the proposed (a) representation and (b) recognition stages.

are characterized by HOGHOF descriptors as in [9] which are used in a Bag of Words framework. For classification we suggest the construction of a vocabulary tree using Hierarchical K-means instead of simple K-means that is used in most approaches. Vocabulary trees (HK-means) have been recently used successfully in object recognition [10] and extended to action recognition (kd-trees) with good results [5]. Finally, we perform classification using an SVM technique with a kernel derived from the difference matrix between the training and testing videos.

A block diagram of the proposed representation method is shown in Fig. 1. The optical flow of the video is initially estimated and used to compensate for camera motion. Motion boundaries are extracted from the motion compensated data and used to extract regions of varying motion. Multi-scale interest points are defined in these areas and tracked with a KLT tracker, whose results are improved by applying RANSAC to the extracted matches. HOGHOF features are estimated over the tracked interest points and used to describe the actions in the video under examination. For recognition, a vocabulary is constructed and used within an SVM framework, as described above.

This paper is organized as follows: Section III presents the motion estimation and tracking process on a dense multi-scale grid in the video. The sequential change detection method that is applied to detect changes in the trajectory points' motion is described in Sec. IV. The HK-means tree based vocabulary construction and subsequent application of SVMs on the matrices of similarity scores are presented in Sec. V. Experimental results and comparisons with existing state of the art methods on new and challenging action recognition benchmark datasets are presented in Sec. VI, while conclusions are drawn in Sec. VII.

III. Dense Multi-Scale Tracking

Motion information is used in conjunction with appearance information to represent actions taking place in a scene. In practice, many scenes are filmed with a moving camera, which renders motion-related features like the Histograms of Optical Flow (HOFs) used here, inaccurate. For this reason, we apply motion compensation to each video's optical flow [7] before any further processing, using the approach of [11] which is based on a bilinear motion model and least squares estimation. In the resulting motion compensated video, gradients of the compensated optical flow are estimated, forming motion boundaries which show when the motion changes. When there is no change in motion, the non-zero motion boundary values are induced by noise, corresponding to hypothesis H_0 below, and are therefore assumed to follow a Gaussian distribution, while changes in motion introduce deviations from Gaussianity (hypothesis H_1).

$$H_0 \quad : \quad u_k^0(r) = z_k(r) \tag{1}$$
$$H_1 \quad : \quad u_k^1(r) = u_k(r) + z_k(r), \tag{2}$$

where $u_k(r), z_k(r)$ denote true and noisy motion-gradient values respectively. The kurtosis G_2 of Gaussian data is equal to zero and can be used to detect whether motion boundaries are caused by noise or by changes in optical flow, in order to localize regions of changing motion. The unbiased estimator of [12] for kurtosis G_2 is given by:

$$\begin{aligned} G_2[y] \quad &= \tfrac{3}{W(W-1)} \textstyle\sum_{i=1}^{W} (u_i(r)^4) - \\ &- \tfrac{W+2}{W(W-1)} \left(\textstyle\sum_{i=1}^{W} u_i(r)^2 \right)^2, \end{aligned} \tag{3}$$

where W is the temporal window within which data (motion values) is obtained. It has been shown that the size of the temporal window does not significantly affect the results [13], and in this work $W = 4$ was used, leading to accurate localization of active pixels. The kurtosis values will be significantly higher in regions of pixels whose motion is changing, so the binarized 2D kurtosis "maps" of each frame, called Motion Boundary Activity Areas (MBAAs) indicate which pixels are undergoing changing motion.

Dense grid points on four scales k_σ (every $k_\sigma = 8, 16, 24, 32$ pixels), hitherto referred to as interest points, are sampled in each frame's MBAA on multiple scales and tracked by using the KLT pyramid tracker [14]. In order to eliminate noisy point matches, RANSAC is applied to the tracked points, retaining only coherent matches for accurate trajectories. HOGHOF descriptors are calculated in multi-scale squares of size $k_\sigma + 8$ around each tracked point to characterize its appearance and motion, as well as determine the optimal trajectory length, as described in Sec. IV.

IV. Real-Time Temporal Segmentation

Once the HOGHOF descriptors over multi-scale tracked interest points are extracted, changes in motion and appearance are detected in order to separate the trajectories in a meaningful manner, related to the actual activity taking place. Changes are found by applying sequential change detection, and in particular the Cumulative Sum (CUSUM) method to the data (HOGHOF feature vectors), as it has been found to provide quickest detection of changes [15]. The HOGHOFs of the first w_0 frames $H_0 = \{\bar{h}_1, \bar{h}_2, ..., \bar{h}_{w_0}\}$ are considered to follow an initial distribution f_0, approximated as a multi-variate Gaussian, with mean and covariance matrix respectively given

by:

$$\bar{\mu}_0 = \sum_{i=1}^{w_0} \bar{h}_i,\ C_0(i,j) = E[(\bar{h}_i - \bar{\mu}_0)^T \cdot (\bar{h}_i - \bar{\mu}_0)] = \\ = \frac{1}{w_0} \sum_{i=1}^{w_0} (\bar{h}_i - \bar{\mu}_0)^T \cdot (\bar{h}_i - \bar{\mu}_0). \quad (4)$$

For simplicity, it is assumed that the HOGHOFs are uncorrelated between different time instances $i \neq j$, giving $C_0(i,j) = 0$ and $C_0(i,i) = \sigma_{0,i}^2$. The resulting initial pdf is given by:

$$f_0(\bar{h}_i) = \frac{1}{(2\pi)^{N/2}|C_0|^{1/2}} \\ \exp\left(-\frac{1}{2}(\bar{h}_i - \bar{\mu}_0)^T C_0^{-1}(\bar{h}_i - \bar{\mu}_0)\right) \quad (5)$$

and at each frame k, it is compared with the "current" pdf, estimated from the w_0 frames before, and including, the current one, i.e. frames $k - w_0 + 1$ to k.

$$f_1(\bar{h}_i) = \frac{1}{(2\pi)^{N/2}|C_1|^{1/2}} \\ \exp\left(-\frac{1}{2}(\bar{h}_i - \bar{\mu}_1)^T C_1^{-1}(\bar{h}_i - \bar{\mu}_1)\right), \quad (6)$$

where $\bar{\mu}_1 = \frac{1}{w_0} \sum_{i=k-w_0+1}^{k} \bar{h}_i$ and $C_1(i,j) = 0$ for $i \neq j$, $C_1(i,i) = E[(\bar{h}_i - \bar{\mu}_1)^T(\bar{h}_i - \bar{\mu}_1)] = \sigma_{1,i}^2$. In order to determine if at a specific frame k these distributions are the same or have changed, the log-likelihood ratio can be used as a test statistic T_k, to be incorporated into the CUSUM test, so we have:

$$T_k = \log\left(\frac{f_1(\bar{h}_k)}{f_0(\bar{h}_k)}\right). \quad (7)$$

The CUSUM test is given by [16] the iterative form:

$$S_k = \max(0, S_{k-1} + T_k),\ S_0 = 0, \quad (8)$$

For Gaussian data under each hypothesis H_0 and H_1, the test statistic, i.e. the log-likelihood ratio, becomes:

$$T_k = \frac{1}{2} \ln\left(\frac{|C_0|}{|C_1|}\right) + \\ + \frac{1}{2}(\bar{h}_k - \bar{\mu}_0)^T C_0^{-1}(\bar{h}_k - \bar{\mu}_0) \\ - \frac{1}{2}(\bar{h}_k - \bar{\mu}_1)^T C_1^{-1}(\bar{h}_k - \bar{\mu}_1), \quad (9)$$

with diagonal covariance matrices given by:

$$C_i = diag[\sigma_{i,1}^2, \sigma_{i,2}^2, ..., \sigma_{i,N}^2],\ i = \{0,1\} \quad (10)$$

where $\sigma_{i,k}^2 = E[(\bar{h}_k - \bar{\mu}_i)^T(\bar{h}_k - \bar{\mu}_i)],\ k = [1,2,...,N]$. The inverse of each diagonal matrix is given by:

$$C_i^{-1} = diag[1/\sigma_{i,1}^2, 1/\sigma_{i,2}^2, ..., 1/\sigma_{i,N}^2] \quad (11)$$

and the determinant of each diagonal matrix is given by $|C_i| = \prod_{j=1}^{N} \sigma_{i,j}^2,\ i = \{0,1\}$.

By plugging in T_k of Eq. (9) into Eq. (8) at each frame, we get a value for the test statistic which significantly increases when there is a change in our data, i.e. the motion/appearance features (HOGHOFs). Thus, at each frame, the value S_k is compared against an empirically derived threshold which is chosen to lead to the lowest number of false alarms and, when it surpasses it, a change is detected. This leads to the temporal segmentation of the extracted trajectories based on actual changes in the activities (motion or appearance) taking place, rather than their segmentation using a manually selected constant threshold.

Fig. 2. Changes detected with CUSUM applied on HOGHOF descriptors respectively. Green lines are tracked trajectories and red lines show where a change has occurred.

Fig. 2 shows some examples of changes detected in the KLT-extracted trajectories based on the HOGHOF descriptors, where the original trajectories are marked in green and become red after a change takes place. In many cases, a change in motion is accompanied by a change in appearance so changes are found in approximately the same frames, albeit in varying regions of the scene.

V. SVM-BASED CLASSIFICATION WITH FAST SEARCH TREES

After the trajectory length N_t for each cuboid is detected, multi-scale spatiotemporal cuboids are created, characterized by the previously extracted HOGHOF features. The $N_c \times N_c \times N_t$ grids around each interest point are divided into $n_\sigma \times n_\sigma \times n_\tau$ cuboids with $n_\sigma = 2$, $n_\tau = 3$, so e.g. a $16 \times 16 \times 15$ cuboid will be split into twelve $8 \times 8 \times 5$ spatiotemporal (ST) cuboids. This cuboid size was chosen as it was found to provide the best average accuracy in our experiments. The HOGHOF features in these smaller cuboids are then used to describe each action by building the corresponding vocabulary. Inspired by [10], we employ Hierarchical K-means (HK-means) to build the vocabulary instead of the commonly used K-means. The reasons for this are that:

1) HK-means allows the use of a much larger vocabulary, which leads to significantly better object recognition results, and is similarly expected to lead to improved action recognition
2) HK-means describes the feature vectors with a hierarchical vocabulary tree, which allows for much faster search and has led to sub-second retrieval times in [10] for a database of a million images.

Thus, a vocabulary is built for all training and testing videos using HK-means. In order to perform classification, we calculate a similarity score between the words in all training and testing videos, $||\bar{q}||$ and $||\bar{d}||$ respectively, based on the L_p norm, as noted in [10]:

$$||\bar{q}-\bar{d}||^p = \sum_i |q_i - d_i|^p = 2 + \sum_{i:q_i\neq 0, d_i \neq 0} (|q_i - d_i|^p - |q_i|^p - |d_i|^p). \quad (12)$$

A pairwise distance matrix $D(\bar{q},\bar{d}) = ||\bar{q}-\bar{q}||^p$ is created from the training data and used to build the kernel $K(\bar{q},\bar{d}) = \exp(-D(\bar{q},\bar{d}))$ which is incorporated into an SVM for classifying test data.

VI. Experiments

Experiments took place on the KTH dataset (www.nada.kth.se/cvap/actions/), as it contains people performing various actions like boxing, running, clapping, etc. The University of Rochester Activities of Daily Living Dataset (www.cs.rochester.edu/ rmessing/uradl/) was also examined, as it contains characteristic activities that people perform in their daily lives and are of interest in many contexts. The Hollywood Action Dataset (HOHA), which is particularly challenging, was also examined in order to test the algorithm's performance in difficult circumstance, and it was shown that it still leads to good results, in some cases even better than the State of the Art (SoA). The accuracy on each dataset was calculated with and without change detection applied on the HOG and HOF descriptors, in order to determine which leads to better activity recognition results.

For the KTH dataset we see that overall good recognition results are achieved, and that the application of change detection on the HOG descriptors leads to improved recognition results. This implies that in these videos, changes in appearance are meaningful for recognizing the type of activity taking place. Table I shows how the average accuracy is affected by varying vocabulary sizes: as expected, there is an improvement for larger vocabularies. The changes in the average accuracy with vocabulary size for the three approaches examined are shown in Fig. 3. Table II shows the average accuracy for each action examined in the KTH dataset, where it can be seen that accurate recognition results are achieved.

TABLE I
AVERAGE ACCURACY FOR KTH WITH VARIOUS VOCABULARY SIZES(%)

Vocab. size	8^4	9^4	8^5	9^5	10^5	10^6
HOGHOF	80.09	82.06	82.06	81.83	**83.45**	83.33
ChDetHOF	82.18	82.64	84.03	82.64	82.99	**84.72**
ChDetHOG	82.87	82.41	83.91	83.91	84.49	**85.53**

For the HOHA dataset it is seen in Table III that we obtain better results than [17] for certain vocabulary sizes and our overall best result is better than the SoA: for a vocabulary size of 9^4, change detection applied on the HOF descriptor gives the best accuracy, namely 30.82%, while the SoA is

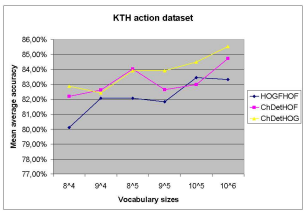

Fig. 3. Effect of vocabulary size on average accuracy for the KTH dataset.

TABLE II
AVERAGE ACCURACY FOR KTH WITH AND WITHOUT CHANGE DETECTION (**85.53%**)

Videos	box	clap	wave	jog	run	walk
box	**100**	0	0	0	0	0
clap	13.89	**85.42**	0.69	0	0	0
wave	15.97	2.78	**81.25**	0	0	0
jog	0	0	0	**80.56**	8.33	11.11
run	0	0	0	30.56	**66.67**	2.78
walk	0	0	0	0.69	0	**99.31**

27.20% on a vocabulary size of 9^5. Fig. 5 depicts the effect of vocabulary size on the HOHA dataset accuracy results as well.

TABLE III
AVERAGE ACCURACY (%) FOR HOHA DATA WITH VARIOUS VOCABULARY SIZES

Voc. size	HOGHOF	ChDetHOF	ChDetHOG	Wang
6^4	27.58	**27.74**	25.83	**24.74**
7^4	**28.14**	27.93	26.78	25.31
8^4	29.36	**30.55**	28.75	27.05
9^4	28.41	***30.82***	29.50	24.69
8^5	**26.84**	26.61	26.90	25.97
9^5	25.18	25.23	25.68	**27.20**
10^5	24.26	24.73	24.08	**26.59**

An analytical presentation of the best results on the HOHA data with our method applying change detection on the HOF descriptor is provided in Table IV:

For the URADL dataset, the application of change detection to either the HOG or HOF descriptor leads to a slight degradation of the results. This can be attributed to the small number of videos used for recognition in these experiments, as well as the nature of the videos: there are not sufficiently significant changes in appearance or motion in each action's "sub-activities" whose detection would improve recognition results. Additionally, it is likely that the application of change

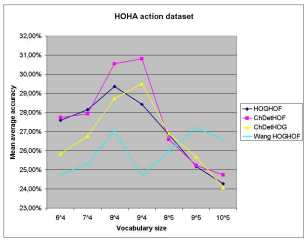

Fig. 4. Effect of vocabulary size on average accuracy for the HOHA dataset.

TABLE IV
AVERAGE ACCURACY (%) FOR HOHA WITH CHANGE DETECTION APPLIED ON THE HOF

HOHA Action	ChDetHOF
Answer phone	17.76
GetOutCar	28.81
HandShake	35.31
HugPerson	33.62
Kiss	44.20
SitDown	42.75
SitUp	6.20
StandUp	37.98
Average	**30.82**

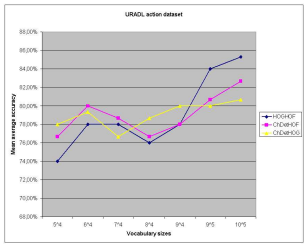

Fig. 5. Effect of vocabulary size on average accuracy for the URADL dataset.

detection to the combined HOGHOF descriptor could provide more meaningful results and better recognition, however this remains an area of future work.

Table V shows how the average accuracy improves for larger vocabulary sizes and Table VI analytically presents the accuracy for each action that is examined. For reasons of space, in the table below the type of activity is represented by a letter, namely: A corresponds to Answer Phone, B to Chop Banana, C to Dial Phone, D to Drink Water, E to Eat Banana, F to Eat Snack, G to Lookup in Phone Book, H to Peel Banana, I to Use Silverware and J to Write on whiteboard. As before, the effect of vocabulary size on the average accuracy is shown in Fig. **??**.

TABLE V
AVERAGE ACCURACY FOR URADL FOR VARIOUS VOCABULARY SIZES (%)

Vocab. size	5^4	6^4	7^4	8^4	9^4	9^5	10^5
HOGHOF	74	78	78	76	78	84	**85.33**
ChDetHOF	76.67	80	78.67	76.67	78	80.67	**82.67**
ChDetHOG	78	79.33	76.671	78.67	80	80	**80.67**

VII. CONCLUSIONS

In this paper we presented an innovative algorithm for human action recognition via dense tracking on multiple scales and estimation of multi-scale HOGHOF descriptors on the tracked points to extract a complete description of the video. RANSAC is applied to the trajectories to eliminate outliers and provide more accurate results. The temporal length of the trajectories is estimated in a novel manner by applying CUSUM sequential change detection to the data in order to find changes in its appearance or motion, which lead to meaningful segmentation of the actions into sub-activities.

TABLE VI
URADL WITH THE HOGHOF FEATURE VECTOR: AVERAGE ACCURACY 85.33%

	A	B	C	D	E	F	G	H	I	J
A	**0.67**	0	0.26	0	0.07	0	0	0	0	0
B	0	**0.93**	0.07	0	0	0	0	0	0	0
C	0.27	0	**0.73**	0	03	0	0	0	0	0
D	0	0	0	**1**	0	0	0	0	0	0
E	0	0	0.07	0	**0.8**	0.13	0	0	0	0
F	0	0	0	0	0	**1**	0	0	0	0
G	0	0	0	0	0	0	**1**	0	0	0
H	0	0.07	0	0.07	0	0.13	0	**0.73**	0	0
I	0	0	0	0.13	0	0	0	0.2	**0.67**	0
J	0	0	0	0	0	0	0	0	0	**1**

Hierarchical K-means is used to construct vocabularies for the actions in the videos in a manner that allows for fast search and leads to very efficient queries, thus allowing the use of a large number of samples, which leads to more accurate results. Experiments have been carried out on challenging benchmark datasets, namely the KTH and URADL videos of human activities, as well as the Hollywood Action dataset HOHA. Good recognition results were obtained in all cases, with improvements for the KTH and HOHA datasets introduced through the application of CUSUM change detection. Future work includes the application of CUSUM change detection on the HOGHOF feature vector and extensive study of the possible feature and codeword combinations for even better recognition of human daily activities.

VIII. ACKNOWLEDGEMENT

This work was funded by the European Commission under the 7th Framework Program (FP7 2007-2013), grant agreement 288199 Dem@Care.

REFERENCES

[1] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, December 2007.

[2] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2002.

[3] H. Wang, A. Klaser, C. Schmid, and C. Liu, "Action recognition by dense trajectories," in *IEEE Conference on Computer Vision & Pattern Recognition (CVPR'11)*, 2011, pp. 3169–3176.

[4] A. Gilbert, J. Illingworth, and R. Bowden, "Action recognition using mined hierarchical compound features," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 5, pp. 883 –897, may 2011.

[5] K. Mikolajczyk and H. Uemura, "Action recognition with appearance-motion features and fast search trees," *Comput. Vis. Image Underst.*, vol. 115, no. 3, pp. 426–438, Mar. 2011.

[6] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Proceedings of the 13th Scandinavian conference on Image analysis*, ser. SCIA'03, 2003, pp. 363–370.

[7] A. Bruhn, J. Weickert, C. Feddern, T. Kohlberger, and C. Schnorr, "Variational optical flow computation in real time," *IEEE Transactions on Image Processing*, vol. 14, no. 5, pp. 608–615, May 2005.

[8] T. Yu and Y. Zhang, "Retrieval of video clips using global motion information," *Electronics Letters*, vol. 37, no. 14, pp. 893 –895, jul 2001.

[9] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conference on Computer Vision & Pattern Recognition, CVPR'08*, 2008.

[10] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, ser. CVPR '06, Washington, DC, USA, 2006, pp. 2161–2168.

[11] V. Mezaris, I. Kompatsiaris, N. Boulgouris, and M. Strintzis, "Real-time compressed-domain spatiotemporal segmentation and ontologies for video indexing and retrieval," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 14, no. 5, pp. 606 – 621, may 2004.

[12] I. V. Blagouchine and E. Moreau, "Unbiased efficient estimator of the fourth-order cumulant for random zero-mean non-i.i.d. signals: Particular case of ma stochastic process," *IEEE Transactions on Information Theory*, vol. 56, no. 12, pp. 6450 – 6458, 2010.

[13] A. Briassouli and I. Kompatsiaris, "Robust temporal activity templates using higher order statistics," *IEEE Transactions on Image Processing*, vol. 18, no. 12, pp. 2756–2768, 2009.

[14] J. Bouguet, "Image sequence enhancement using multiple motions analysis," in *Intel Corporation. Microprocessor Research Labs*.

[15] V. P. Dragalin, "Optimality of a generalized cusum procedure in quickest detection problem," in *Statistics and Control of Random Processes: Proceedings of the Steklov Institute of Mathematics*, Providence, Rhode Island, 1994, pp. 107–120.

[16] E. S. Page, "Continuous inspection scheme," *Biometrika*, vol. 41, pp. 100–115, 1954.

[17] H. Wang, M. M. Ullah, A. KlŁser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *British Machine Vision Conference*, 2009.