Activity detection and recognition of daily living events

Konstantinos Avgerinakis
Information Technologies Institute,
Centre for Research & Technology
University of Surrey
koafgeri@iti.gr

Alexia Briassouli
Information Technologies Institute,
Centre for Research & Technology
abria@iti.gr

Ioannis Kompatsiaris
Information Technologies Institute,
Centre for Research & Technology
ikom@iti.gr

ABSTRACT

Activity recognition is one of the most active topics within computer vision. Despite its popularity, its application in real life scenarios is limited because many methods are not entirely automated and consume high computational resources for inferring information. In this work, we contribute two novel algorithms: (a) one for automatic video sequence segmentation - elsewhere referred to as activity spotting or activity detection - and (b) a second one for reducing activity representation computational cost. Two Bag-of-Words (BoW) representation schemas were tested for recognition purposes. A set of experiments was performed, both on publicly available datasets of activities of daily living (ADL), but also on our own ADL dataset with both healthy subjects and people with dementia, in realistic, life-like environments that are more challenging than those of benchmark datasets. Our method is shown to provide results better than, or comparable with, the SoA, while we also contribute a realistic ADL dataset to the community.

Categories and Subject Descriptors

I.5.4 [**PATTERN RECOGNITION**]: Applications—*Computer vision*

General Terms

Algorithms, Experimentation

Keywords

Activity detection, Activity recognition

1. INTRODUCTION

In recent years, Ambient Assisted Living (AAL) solutions are being developed to help people with chronic degenerative conditions continue living independently for as long as they can. This is achieved in large part by continuous unobtrusive monitoring for accurate activity, lifestyle and behavioral profiling, which ensures their safety in case of an emergency, as well as the detection of gradual changes in their condition. The results of the monitoring and profiling provided by such systems can then be used as input in appropriate feedback both for the people being monitored, as well as for their carers.

Existing assisted living solutions usually employ physiological and environmental sensors that are relatively simple, like accelerometers and contact sensors. Attention has recently turned to the use of more sophisticated technologies, based on audiovisual monitoring. In this work we focus on the use of video for remote monitoring of, for example, people with conditions like dementia, living home alone. For effective video-based monitoring, the recognition of ADLs is central, and also the focus of this work. Activity recognition from video for assisted living is based on unobtrusive ambient sensors, namely static video cameras, which do not disturb people in their daily life. It is essential to provide highly accurate recognition results, to build useful activity, lifestyle and behavioral patterns for each human subject and help their carer remotely monitor the progress of their condition to help them accordingly.

Activity recognition in computer vision mainly focuses on extracting information from pre-segmented video sequences, which makes it inappropriate for dealing with real scenarios, where videos are not segmented beforehand. In real life scenarios scene conditions are challenging and diverse and near real time results may be required, especially for the detection of emergencies. In practical situations, activity recognition needs to be preceded by activity detection, which localizes an action of potential interest in time in a video.

Early activity recognition analyzed simple, constrained scenarios [8], [23], while more challenging datasets are now coming to the attention of the activity recognition community, featuring camera motion, greater anthropometric variance and changes in scene illumination. The current SoA in activity recognition analyze Hollywood movies [13], [17], sports videos [20], [21] or activities recorded in unconstrained conditions, such as YouTube videos [15]. This has led to the development of more sophisticated algorithms, but at the cost of a high computational burden, which does not allow the deployment of these methods in realistic scenarios. Recently, ADL datasets were introduced in the literature [19], [22], depicting common activities of daily life, performed by several human subjects. Their focus is on real life scenarios, but they do not address the problem of activity detection, which requires the automatic detection of activities in time. Instead, activities are detected manually, and

then classified into one of specific categories under examination. In this work, we adapt activity recognition to the realistic data and needs of an assisted living home environment, where continuous video is available, with no options for manually segmenting it in time. Motivated by these reasons, we propose robust activity detection algorithm and a light activity recognition technique for reliable and fast activity recognition.

The paper is organized as follows: related work is presented in Section 2, while Section 3 describes the proposed activity detection technique. Section 4 describes the activity representation and recognition algorithms used, while experiments are presented in Section 5 proving the robustness of our technique in publicly available videos, as well as in our own dataset. Section 6 completes the work with useful conclusions.

2. RELATED WORK

Despite the high popularity that activity recognition has gained the last years, to the best of our knowledge, only a few works focus on the activity detection problem for automatic video sequence segmentation [6], [11], [14], as attention has mostly been placed on accurately classifying video sequences and not on temporally localizing activities in a video. It should be emphasized that activity detection is different from shot detection, as several activities may take place within a shot, either sequentially or simultaneously. We contribute a novel activity detection technique to automatically segment video sequences based on a simple but robust statistical technique.

Activity recognition, on the other hand, has been thoroughly studied during the last decade and usually consists of two parts, namely activity representation and activity recognition. For activity representation, the literature can be split into holistic approaches, such as motion history volumes [26], space-time shapes [8], trajectory descriptors [19], [25], [18] and temporal templates [2], and local based ones that use 3D local patches such as [13], [25], [24], [10], [27], which are either based on the extension of local patches to the temporal space (i.e. SIFT3D, HOG3D, SURF3D in [24], [10], [27]) or on the construction of motion histograms around sampled interest points (i.e. HOF, MBH in [13], [25]). Interest points can be sampled either in a sparse manner, as in [27], [12], [7] or densely [25], the latter providing better recognition rates than the former.

In this work, we densely sample interest points after applying a background subtraction technique based on the higher order statistical analysis of motion in the video sequences. In order to represent activities in videos, we use a local approach, which encodes both appearance and motion information (HOGHOF descriptor), enriched with holistic features extracted from raw trajectory cues. This addition of spatial information in our BoW models is shown to increases recognition rates. Activity recognition based on local features usually combines Kmeans clustering with a Chi-Square kernel, resulting in a BoW representation with hard binning. Inspired by recent State of the Art (SoA) results in image classification [9], [5], we suggest instead a soft binning approach based on Gaussian Mixture Model (GMM) clustering combined with Fisher vectors.

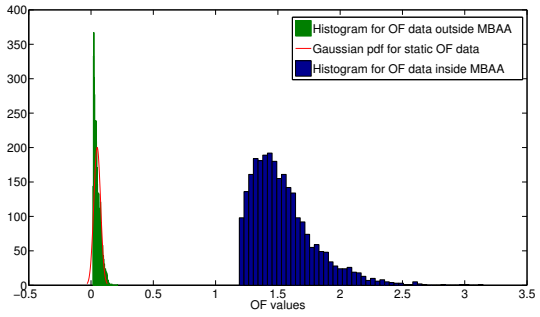

Figure 1: Optical flow histograms for static (green) and moving (blue) pixels. The Gaussian model of our static data is shown in red.

3. ACTIVITY DETECTION

Activity detection localizes video subsequences in time that contain potentially interesting information, such as activities and events to be recognized. These parts of the video are subsequently passed to an activity recognition algorithm. By this technique, un-interesting parts of videos, such as motionless frames or long video subsequences with the same activity, are successfully ignored.

For activity detection, we use motion detection combined with trajectory extraction. We first detect regions of interest (RoI), where motion undergoes changes, and then sample interest points in them and track them over time until the motion/activity of interest ends, producing the video sequence to be processed. Our motion detector relies on kurtosis-based Activity Areas [3], for which optical flow values [4] are analyzed statistically and regions that contain motion throughout successive frames are localized in a real-time manner. The method of [3] separates moving pixels from static ones by considering that the inter-illumination differences of static pixels follow an approximately Gaussian distribution whose kurtosis is nearly zero. In this work, we consider that noise-induced estimates of optical flow of static pixels are approximately Gaussian, as shown in the histograms of aggregated optical flow values in Figure 1, which is acquired from the analysis of 10 sample videos of URADL action dataset [19]. Good separation results among the two classes have also been observed in further action datasets, such as KIT [22], HOHA [13] and UCF [21]. To further verify this, we applied the Kolmogorov-Smirnov test [16] on ten training videos from the URADL action dataset [19], which is often used to determine if a dataset indeed follows a Gaussian distribution. The application of this test led us to conclude that the optical flow of static pixels is adequately modeled by a Gaussian distribution, unlike that of moving pixels. Thus, Activity Areas are extracted to separate moving pixels from static ones by monitoring optical flow values over successive video frames. The following two hypotheses represent the optical flow estimates $u_t^0(x,y)$ induced by noise (H_0) and actual motion $u_t^1(x,y)$ (H_1):

$$\begin{aligned} H_0 &: u_t^0(x,y) = z_t(x,y) \\ H_1 &: u_t^1(x,y) = u_t(x,y) + z_t(x,y) \end{aligned} \quad (1)$$

where $u_t(x,y)$ denotes the true optical flow at pixel (x,y) at time t and $z_t(x,y)$ the additive noise in that location. A fast and robust way to separate the data following a Gaussian

distribution (H_0) from true optical flow values (H_1) is to estimate the data kurtosis over time and binarize the resulting flow map. A novel technique for accurately estimating the empirical value of the kurtosis [1] in an unbiased manner approximates the excess Kurtosis by the fourth-order cumulant estimator:

$$G_2[y] = \frac{3}{W(W-1)} \sum_{i=1}^{W} \left(u_i(x,y)^4 \right) - \frac{W+2}{W(W-1)} \sum_{i=1}^{W} \left(u_i(x,y)^2 \right)^2 \quad (2)$$

where W is the temporal window over which Kurtosis values are computed, set equal to $W = 10$. Further experimentation showed that values in the range $W = 10 - 20$ lead to equally accurate Activity Areas (AAs).

The kurtosis values lead to AAs by binarizing the data as below, suing a threshold acquired empirically from several videos of the URADL action dataset:

$$AA(x,y) = \begin{cases} 0 & \text{if } G_2[x,y] < 2 \cdot 10e^{-2} \\ 1 & \text{else} \end{cases}$$

From these regions, we sample candidate interest points and track them with a boosted KLT tracker until they become motionless, after which the video subsequence is considered to end.

A spatial grid is formed in each video frame's AAs producing a number of blocks where activity detection takes place. The center pixel (x, y) of each block is considered as a candidate interest point only when more than 50% of the block belongs to the moving pixels of the AA. All candidate interest points are tracked using KLT [28], boosted by a homography test, which uses a RANSAC estimator to validate interest point correspondences. Interest points that pass the test are used to track moving objects and form a trajectory descriptor, further analyzed in Section 3. When no more interest points exist in the scene, the video subsequence is considered to end and activity recognition takes place, so as to detect the activity taking place in those frames. Figure 2 depicts an example where activity detection localizes the start and the end frame of an activity subsequence. An in between frame is also depicted for presentation purposes.

Figure 2: Activity detection throughout a video. Activity Areas are depicted on the left of each frame while on the right trajectories are depicted. Green trajectories are active, while red are the terminated ones.

4. ACTIVITY RECOGNITION

Activity recognition algorithms comprise of: (a) activity representation and (b) activity recognition, which are highly time consuming, as they need to address difficult problems, such as camera motion, scale variations, illumination changes, often present in real, unconstrained environments. In the case of monitoring to support independent living at home, indoors activities are recorded by static cameras, while scene illumination is relatively stable, computational cost, both for representation and recognition, allowing a real time scenario to be implemented.

4.1 Activity representation

In order to benefit from the advantages of both local and global state-of-the-art activity recognition techniques, we have concluded that a hybrid descriptor can lead to increased recognition rates at a lower computational cost.

Thus, we use a local approach for describing appearance and motion characteristics of our activity and a holistic one for global spatial information. Sampled interest points, already extracted for activity detection, provide us with very accurate trajectory vectors. Regions around interest points are described in four spatial scales for scale invariance in our activity descriptor. We use one of the fastest methodologies for extracting the activity descriptor, called HOG-HOF proposed in [13]. HOG histograms sustain appearance information, while HOF provides motion characteristics. The spatiotemporal descriptor is formed by concatenating all histograms that belong to the same trajectory. Each descriptor is subdivided into a $(n_x = n_y = 2, n_t = 3)$ grid of cuboids and, for each cuboid, coarse histograms are averaged and normalized to extract rich information in each volume. Raw trajectory coordinates are added to the vector to include global spatial information to our descriptor.

Let $HOG(B_{sc}, t)$ and $HOF(B_{sc}, t)$ be the histograms extracted inside a trajectory block $B_{sc} = (x, y, w_{sc}, h_{sc})$, around an interest point with coordinates (x, y) and size :

$$(w_{sc}, h_{sc}) = \left(\sum_{i=0}^{3} 8 + i \cdot 8, \sum_{i=0}^{3} 8 + i \cdot 8 \right),$$

where w_{sc} and h_{sc} are the width and height respectively. The index sc denotes the different sizes that the block might take, depending on the scale size. The resulting spatiotemporal descriptor around each interest point (x, y) is the L_2 normalized concatenation of the averaged histograms within each temporal sub-volume that is formed by dividing the initial descriptor by N/n_t :

$$ST_{desc} = \Bigg[concat_{j=1}^{n_t} \Bigg\{ \Bigg\| \sum_{t=(j-1)\cdot(N/n_t)+1}^{j \cdot N/n_t} \frac{HOG(B_{sc}, t)}{N/n_t} \Bigg\|_2 \ldots \Bigg\| \sum_{t=(j-1)\cdot(/n_t)+1}^{j \cdot N/n_t} \frac{HOF(B_{sc}, t)}{N/n_t} \Bigg\|_2 \Bigg\} (x_t, y_t) \Bigg]$$

where ST_{desc} denotes the final feature vector formed by

the concatenation, denoted here as *concat*, of the spatio-temporal volumes and is used for the representation of each activity. $HOG(B_{sc}, t_i)$ and $HOF(B_{sc}, t_i)$ are both represented below as *hist*. Each block histogram consists of $n_x \times n_y$ cells. These cells vote for the construction of the final block histogram and are computed as seen bellow:

$$hist_{block}(B_{sc}, t_i) = \Big\| accumulate_{i,j} \Big(hist(x + i \cdot \frac{w_{sc}}{n_x}, ... \\ , y + j \cdot \frac{w_{sc}}{n_y}, \frac{w_{sc}}{n_x}, \frac{h_{sc}}{n_y}) \Big) \Big\|_2$$

where $(i, j) = \{(1,1)(1,-1)(-1,1)(-1,-1)\}$ and *hist* returns the spatial block histogram around each trajectory interest point. Consequently, $hist_{block}$ is the accumulation, here denoted as *accumulate*, of its 4 cell histograms.

4.2 Activity recognition

For activity recognition we tested and compared two different methods. The first uses K-means and a Chi-Square Kernel. In this recognition schema, cluster centers are extracted using K-means and Bag-of-Words follows a hard binning approach to create histograms of visual-words. Chi-Square is used for creating a distance kernel among them and is fed to an SVM classifier to characterize activities in testing videos. In our experiments, we use K=4000 cluster centers to partition the feature vector space. To limit complexity, cluster centers are clustered on a randomly selected subset of 100.000 feature vectors acquired from the training set. K-means is initialized 10 times in order to provide the most discriminative cluster centers.

The second technique that was tested is inspired from recent good results in image classification [9], [5]: GMM is used to define vocabulary cluster centers, while soft-binning is used to create a Bag-of-words representation for each video. Fisher vector distances between the activity descriptor and cluster centers create an analytic description for each video. In our experiments we tested a much smaller vocabulary size (K=256) than that used with K-means, which proved to work more quickly and also result in higher accuracy than that obtained with larger K-means vocabularies.

5. EXPERIMENTS

Experiments took place on publicly available ADL datasets to determine the applicability and robustness of our algorithm. Testing our algorithm on such a large number of videos also gave us the opportunity to detect limitations or omissions of current ADL datasets: there is insufficient anthropometric variance, while environmental conditions are constrained, making current benchmark data in appropriate for testing real-life situations. Thus, we recorded ADL data from the Greek Association of Alzheimer's Disease and Related Disorders (GAADRD), both from healthy individuals and from people with mild dementia to Alzheimer's (AD) performing ADLs in a home-like environment.

5.1 URADL dataset

URADL [19] is a well-known dataset for recognizing ADLs and is used mostly for evaluating purposes. In this dataset, 5 different actors were called to perform 10 different activities, 3 times in a kitchen environment. A serious disadvantage of this dataset is that it lacks anthropometric variance, while environmental conditions are quite simple, since the activities take place in the same location, with the same illumination, a static camera, no environmental noise and no occlusions. For evaluating our algorithm, we choose to use leave-one-subject-out testing. Thus, we initialized the recognition procedure 5 times, so that we can recognize the activity of each human subject independently. The names of the activities of this dataset are encoded in the tables as AP = Answer Phone, CB = Chop Banana, ES = Eat Snack, DP = Dial Phone, DW = Drink Water, EB = Eat Banana, LiP = Look up in Phonebook, PB = Peel Banana, US = Use Silverware, WoW = Write on Whiteboard. Figure 3, depicts some characteristic URADL activities grabbed from some video frame samples.

Figure 3: Characteristic video frames for answer phone, drink water and eat banana activities, taken from URADL dataset from left to right.

Table 1: Aggregated results for URADL

	HOGHOF without RANSAC	HOGHOF with RANSAC	Vel.Hist [19]	MBH [25]
Kmeans & ChiSquare	87,3%	89,3%	89,3%	89,3%
GMM & Fisher	90,0%	90,6%	n/a	n/a

It is obvious from Table 1, that GMM & Fisher recognition performs better rather K-means & Chi-Square. This table also shows that RANSAC helped correct some erroneous correspondences in the trajectory structure. Our best result, seen in Table 2, is better than those of both SoA activity recognition methods [19, 25]. From Table 2, we observe that our algorithm leads to accurate activity recognition: only "answer phone" was recognized with low accuracy, as it was confused with "dialing phone".

5.2 Dem@Care ADL dataset

Taking into account the pros and cons of current public ADL datasets, we proceed with the launch of a new set of recordings, held in GAARD premises in Thessaloniki. Several people participated in the experiments, including people with dementia, people with mild cognitive impairment (MCI) and healthy ones. Many people were tested (32 people) introducing great anthropometric variations in our activity dataset, while the videos contain various activities. The human subjects were called to perform a set of activities encoded as follows: CU: clean up table, DB: drink beverage (i.e. water-orange juice), EP: end phone-call, ER: enter room, ES: eat snack, HS: handshake, PS: prepare snack, RP: read paper on the couch, SB: serve beverage, SP: start phone-call, TV: talk to visitor. For evaluating our algorithm, we choose to use two different splits of the dataset shown in Table 3 and Table 4. The first experimental setup

Table 2: Our best recognition result on URADL action dataset, when HOGHOF representation was combined with GMM & Fisher recognition schema.

	AP	CB	DP	DW	EB	ES	LiP	PB	US	WoW
AP	46,7%		33,3%		20%					
CB		93,3%						6,7%		
DP	6,7%		93,3%							
DW				100%						
EB			6,7%		80%	6,7%		6,7%		
ES						100%				
LiP							100%			
PB						6,7%		93,3%		
US									100%	
WoW										100%
Av.Acc	90,7%									

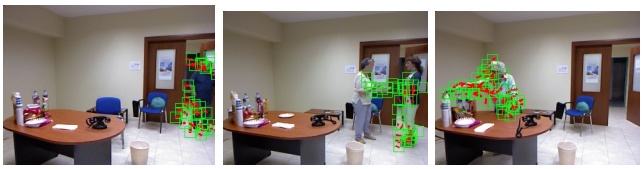

Figure 4: Characteristic video frames for enter room, handshake and serve beverage activities, taken from URADL dataset from left to right.

follows a leave-one-subject-out splitting of the dataset, while in the second we separate 20 subjects for training and keep 12 subjects for testing. Figure 4, depicts some characteristic Dem@Care activities grabbed from our videos.

Table 3: 20 train to 12 test video samples recognition results in Dem@Care ADL dataset.

	HOGHOF without coords	HOGHOF with coords	HOGHOF [25]	MBH [25]
Kmeans & ChiSquare	80,4%	74,6%	84,2%	77,4%
GMM & Fisher	79,4%	82,5%	80,9%	77,3%

Table 4: Leave-One-Subejct-Out recognition results in Dem@Care ADL dataset.

	HOGHOF without coords	HOGHOF with coords	HOGHOF [25]	MBH [25]
Kmeans & ChiSquare	85,9%	83,9%	93,5%	93,7%
GMM & Fisher	86,2%	91,2%	94,1%	93,4%

When using the 20 training - 12 testing video split of the Dem@Care dataset, seen in Table 3, we surpass the SoA in most cases. The inclusion of raw trajectory coordinate data boosts the representation when it is combined with GMM & Fisher recognition schema and performs far better the K-means & Chi-square method. In the second experimental setup, where leave-one-subject-out was followed, our results were comparable to the SoA, while our recognition rates improved significantly. Finally, Table 5 shows that our method accurately classifies most activities in the Dem@Care data, despite the great anthropometrics variance and realistic conditions depicted in these videos. Furthermore, we can observe from the experiments a peculiar behaviour on the performance of MBH descriptor. Contrary to HOGHOF action descriptor [25], MBH recognition rates seems to drastically differentiate among the two experiments. This can be explained by the fact that MBH are based on the analysis of actions that occur on the boundary of the human patient who performs the action. Thus, as the camera distance from the human patient increase, so does these regions become smaller, producing less discriminative representations and consequently lower recognition rates. However, as we can observe from the two experiments, if we increase the video samples that are used for training purposes, we can also acquire very accurate recognition rates for MBH, even comparable to HOGHOF action descriptor.

6. CONCLUSION

In this work, we introduce a comprehensive solution for activity detection and recognition, where large videos are initially split into video subsequences containing potentially interesting activities, which are then processed in for accurate activity recognition. Activity detection is based on the analysis of motion vectors over time using in a theoretically sound statistically method, rather than heuristics. Activity recognition methods based on the SoA in the fields of both object and activity recognition are then tested on the temporally localized subsequences. Experiments on benchmark datasets and a new, more challenging and realistic dataset recorded at the GAARD in Thessaloniki, Greece, show that our method obtains highly accurate recognition rates, comparable to, or surpassing the SoA, making it appropriate for real life applications.

7. ACKNOWLEDGEMENTS

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

Table 5: Our best result in Dem@Care action dataset when using HOGHOF with raw trajectory coordinates for activity representation and GMM & Fisher recognition schema. Leave-one-subject-out split setup was followed on this case.

	CU	DB	EP	ER	ES	HS	PS	RP	SB	SP	TV
CU	83,8%				10,3%		5,9%				
DB	0,7%	98,3%			1%						
EP	3,1%		89,1%		3,1%					4,7%	
ER				100%							
ES		10,9%	2,2%		86,6%			0,3%			
HS						93,8%					6,3%
PS		2,9%			8,6%		83,4%		5,1%		
RP	3,1%							96,9%			
SB					1,5%		11,8%		86,8%		
SP			6,1%		6,1%					87,9%	
TV						3,2%					96,8%
Av.Acc	91,2%										

8. REFERENCES

[1] I. 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. *Information Theory, IEEE Transactions on*, 56(12):6450–6458, 2010.

[2] A. F. Bobick, J. W. Davis, I. C. Society, and I. C. Society. The recognition of human movement using temporal templates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:257–267, 2001.

[3] A. Briassouli and I. Kompatsiaris. Robust temporal activity templates using higher order statistics. *Image Processing, IEEE Transactions on*, 18(12):2756–2768, 2009.

[4] A. Bruhn, J. Weickert, and C. Schnorr. Lucas/kanade meets horn/schunck: Combining local and global optic flow methods. *International Journal of Computer Vision*, 61:211–231, 2005.

[5] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *British Machine Vision Conference*, 2011.

[6] K. G. Derpanis, M. Sizintsev, K. Cannons, and R. P. Wildes. Efficient action spotting based on a spacetime oriented structure representation. In *In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2010.

[7] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*, pages 65–72, 2005.

[8] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. In *In ICCV*, pages 1395–1402, 2005.

[9] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid. Aggregating local image descriptors into compact codes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(9):1704–1716, 2012.

[10] A. Klaser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *In British Machine Vision Conference*, 2008.

[11] A. Klaser, M. Marszalek, C. Schmid, and A. Zisserman. Human focused action localization in video. In K. N. Kutulakos, editor, *ECCV Workshops (1)*, volume 6553 of *Lecture Notes in Computer Science*, pages 219–233. Springer, 2010.

[12] I. Laptev and T. Lindeberg. Space-time interest points. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 432–439 vol.1, 2003.

[13] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Conference on Computer Vision & Pattern Recognition*, jun 2008.

[14] I. Laptev and P. Perez. Retrieving actions in movies. *IEEE International Conference on Computer Vision*, 2007.

[15] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos "in the wild", 2009.

[16] G. Marsaglia, W. W. Tsang, and J. Wang. Evaluating kolmogorov's distribution. *Journal of Statistical Software*, 8(18):1–4, 11 2003.

[17] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *Conference on Computer Vision & Pattern Recognition*, jun 2009.

[18] P. Matikainen, M. Hebert, and R. Sukthankar. Trajectons: Action recognition through the motion analysis of tracked features. In *Workshop on Video-Oriented Object and Event Classification, ICCV 2009*, September 2009.

[19] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *ICCV '09: Proceedings of the Twelfth IEEE International Conference on Computer Vision*, Washington, DC, USA, 2009. IEEE Computer Society.

[20] J. C. Niebles, C.-W. Chen, , and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *Proceedings of the 12th European Conference of Computer Vision (ECCV)*, Crete, Greece, September 2010.

[21] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach: a spatio-temporal maximum average correlation height filter for action recognition. In *In Proceedings*

of IEEE International Conference on Computer Vision and Pattern Recognition, 2008.

[22] L. Rybok, S. Friedberger, U. D. Hanebeck, and R. Stiefelhagen. The KIT Robo-Kitchen Data set for the Evaluation of View-based Activity Recognition Systems. In *IEEE-RAS International Conference on Humanoid Robots*, 2011.

[23] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local svm approach. In *In Proc. ICPR*, pages 32–36, 2004.

[24] P. Scovanner, S. Ali, and M. Shah. A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th international conference on Multimedia*, MULTIMEDIA '07, pages 357–360, New York, NY, USA, 2007. ACM.

[25] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action Recognition by Dense Trajectories. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 3169–3176, Colorado Springs, United States, June 2011.

[26] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. In *COMPUTER VISION AND IMAGE UNDERSTANDING*, 2006.

[27] G. Willems, T. Tuytelaars, and L. Gool. An efficient dense and scale-invariant spatio-temporal interest point detector. In *Proceedings of the 10th European Conference on Computer Vision: Part II*, ECCV '08, pages 650–663, Berlin, Heidelberg, 2008.

[28] J. yves Bouguet. Pyramidal implementation of the lucas kanade feature tracker. *Intel Corporation, Microprocessor Research Labs*, 2000.