Video Segmentation Descriptors for Event Recognition

Remi Trichet
Inst. Robotics & Intelligent Systems
USC, Los Angeles, CA 90089-0273
Remi.trichet@gmail.com

Ramakant Nevatia
Inst. Robotics & Intelligent Systems
USC, Los Angeles, CA 90089-0273
nevatia@usc.edu

***Abstract*—This paper presents a new video motion descriptor based on a multi-scale video segmentation to provide a multi-layered output as well as connections with the rich interactions that occur between objects at the semantic level. We also put the emphasis on relationships between motion clusters by providing a new relative motion descriptor encapsulating relative motion patterns within a local spatio-temporal neighborhood. Experimental results on the challenging TRECVID MED11 event recognition dataset validate the approach.**

I. Introduction

Local features are a common means to represent visual media. Their availability as well as their outstanding performance when associated with the Bag-of-Words technique, have made them a popular choice. Over the past decade, they have been extended from images to videos and applied to a wide range of related tasks, such as activity recognition or video indexing.

Looking back at the myriad of methods that have been developed so far, several lessons have been learned:

First, 3D appearance patches [1]–[6] are not sufficient to characterize videos. As the experience [7]–[10] has proved, motion patterns fine representation is paramount. These descriptors, typically grounded on tracks or tracklets, need to be robust to camera motion, noise and the data variability [11], [16].

Second, pairwise relationships representation of moving volumes [9], [12], [13], [17] improves performance. These relationships can be spatial and/or temporal and require a video decomposition or clustering in homogeneously moving areas.

Third, modeling background (sometimes referred as global context descriptor) separately from the moving objects [18], [20]–[22], [30] leads to a richer representation, depicting actions as well as the scenery in which they take place.

Finally, dense sampling has recently proved to outperform sparse sampling on videos for activity recognition by Wang *et al.* [15]. More recently, the same experiment was applied to trajectories [10], showing that a dense set significantly surpasses a sparse one. However, this strategy implies extensive feature computation, which may not always be practically feasible for large datasets.

Despite all these progress, several directions remain to be explored. In this paper, we tackle two of them. First, we argue that current techniques lack some motion pattern scalability as codeword proximity definition is not necessarily robust to the temporal and spatial granularity of the action area they lie into.

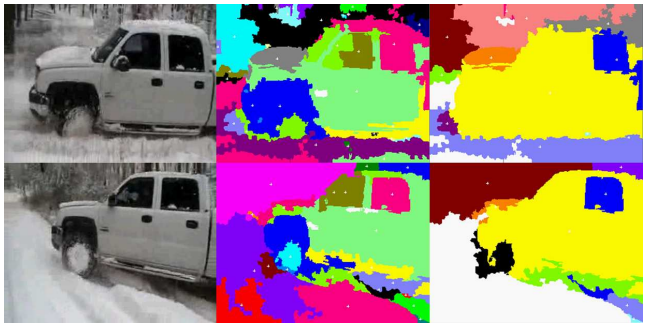

Fig. 1: Spatio-temporal tube segmentation example. The first column displays original frames *f* and *f+50*. Last two columns show the segmentation at the low and high scales.

Our second motivation is based on the comparison of typical statistical approaches with structural approaches that detect objects of interest, analyze their appearance and motion patterns over time to finally infer a conclusion. If the former outperforms the latter, it is at the cost of semantic information. Indeed, in the case of a successful recognition of an event (like 'fight') within a given video clip by the first technique, the algorithm will not be able to give any extra descriptive information, like at what frame and spatial location of the video the event occurs, and which atomic actions (such as 'punch', 'chase', 'kick') have been leading to that conclusion. We aim to develop a video descriptor that achieves state-of-the-art performance while preserving video semantic information.

To address these issues, we propose to ground video motion and appearance descriptors on a video segmentation producing a hierarchical, multi-scale output. An example is provided in figure 1. This layout naturally embeds the video structure, and therefore yields robust and scalable descriptors. To the best of our knowledge, this is the first descriptor based on such mid-level features which makes it a perfect tool for tasks such as human detection or discriminative patch selection.

The segmentation utilizes optical flow to track superpixels over time. These volumetric tracks, that we call spatio-temporal tubes (STTs) are then fused according to their motion and color patterns. As the full fusion tree is stored, the output is a hierarchy of volumes, dubbed *supertubes*. Each

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.339

supertube is then broken in overlapping temporal volumes that are represented with color, HoG3D and relative motion histograms (RMH). This last, new histogram encapsulates accurate pairwise motion relationships of a supertube neighborhood. Context is finally modeled by quantizing on spatial area sizes to yield different histograms. Results validate the approach.

The remainder of this paper is organized as follows. Section 2 discusses the related work. Section 3 gives an overview of the method. Section 4 briefly explains the video segmentation on which the descriptors are grounded and section 5 details the motion descriptors. Section 6 demonstrates the effectiveness of the approach through experiments on web videos. Finally we conclude and give a peek at our future work in section 7.

II. Related work

An early attempt in extracting spatio-temporal features was Laptev and Linderberg STIP features [1], adapting the Harris corner detector to the third dimension. Following this first impetus, numerous 3D extensions of effective image descriptors such as HoG3D [2], MoSIFT [3], SURF [4], 3D-SIFT [5], and LTP [6], have emerged.

However, solely describing appearance of volumes neglects the motion patterns that are naturally embedded in videos. To face this problem, approaches tracking interest points have emerged. Messing *et al.* [7] track Harris3D interest point [1] using KLT optical flow [8]. Sun *et al.* [9] approach the issue by quantifying point-wise visual cues, intra-trajectory motion information, and inter-trajectory motion proximity features for each tracklet. Wang *et al.* [10] consider a dense trajectory (DT) set to extract HoGs and HoFs along with nearby boundaries histogram (MBH) making the approach robust to camera changes [11]. Nogushi and Yanai [16] proposed tracking SURF features over time and grouping the tracks falling within the same triangles of a Delaunay mesh. This grouping strategy offers an interesting alternative to the motion approximation over large regions but finding the right triangle size adjusting to the scene topology among the swarm of extracted features remains a problem.

The next step was the introduction of relationships between large, stable, trajectory clusters. Sun *et al.* [9] used simple proximity relationships. Mattikainen *et al.* [17] explored this area modeling spatial and temporal relations separately in a computationally efficient way. Raptis and Soatto [12] proposed a dense tracklet-based descriptor. Due to the limited lifespan of these trajectories, this work was focusing on local structure patterns of trajectories. Jiang *et al.* [13] extended this paradigm by modeling pairwise relationships between long trajectory clusters. However, their reasoning is only performed at the codeword level.

Arguing that the scenery in which an action takes place is often correlated to the action itself, some authors also utilize global context descriptors in addition to the local ones. This idea was successfully applied on SIFTs [19] by [20], [21]. Carmichael *et al.* [22] similarly modeled the global context for SURF [4] and MSER [23]. More comprehensive experiments were performed by Marszalek *et al.* [30]. Ballas *et al.* [18] utilized a deformable adaptive grid to closely adjust to the object boundaries and more efficiently separate them from the background.

However, little effort has been made so far on analyzing motion pattern at various spatio-temporal scales.

III. Outline of the method

The algorithm, depicted in figure 2, proceeds as follow. Tube determination is grounded on the segmentation of each frame in a set of compact and color-wise uniform superpixels (2.b). Optical flow is determined for each frame (2.c) and utilized to determine each superpixels matching candidates within the next frame (2.d). Exhaustive matching is then performed according to the color similarities. Once all the frames have been processed, spatio-temporal tubes are then grouped using a meanshift algorithm [24] (2.e). Finally the set of supertubes is iteratively fused (2.f). The complete set of hierarchically fused supertubes (from fine to coarse) as well as the underlying tree structure is outputted (2.g).

Each supertube is then temporally chunked in overlapping volumes (2.h). Each volume is described with 3 histograms (2.i): a Lab color histogram, a HoG3D histogram and a relative motion histogram (RMH). Each volume histograms are then concatenated to yield the final descriptor. The set of descriptors are quantized according to a n-codewords dictionary (2.j) and each video is represented with 9 histograms of size n according to the different descriptors layout spatial sizes and motion intensities (2.k).

IV. Video segmentation

Various video segmentation techniques have been proposed based on superpixels [26], [28], [31]–[33]. We use [28] that allows a hierarchical output and application to large datasets. This section briefly describes this segmentation process.

A. Video segmentation features

The main components of our segmentation algorithm are superpixels due to their capacity to summarize local information and the low variability of their spatial layout from one frame to another. We further will track them from frame to frame. We chose the SLIC superpixels [25], spatially uniform and stable over time.

Each superpixel is described according to 4-motion and 3-appearance values. As our method performs video segmentation with respect to the motion patterns, motion assessment is a core component. Hence, optical flow is computed according to [27] which offers the advantages of dense motion estimation and readily available code. The L, a, b values of the CIELAB colorimetric system which has proved to be more accurate for computer vision problems, is used as appearance representation.

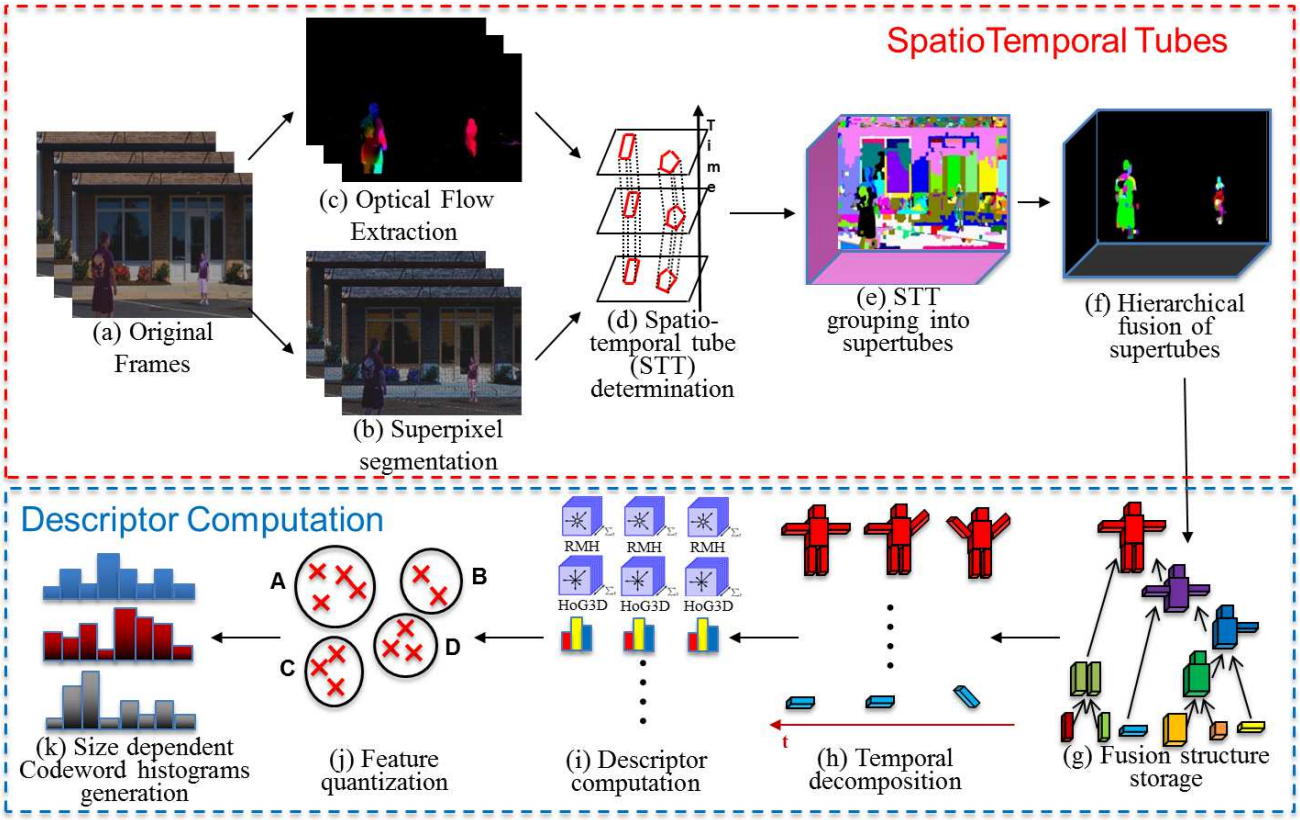

Fig. 2: Outline of the approach. See text for details. Best viewed in color.

B. STT extraction

STTs are generated by matching superpixels from frame to frame. The matching candidates of a given superpixel are determined according to its computed optical flow. Superpixel matching from one frame to another is constrained according to color, and overlap. We also require a significant amount of superpixel pixels to be matched for the superpixel not to be considered occluded.

The algorithm favors long tracks, leaves a lot of freedom for motion and spatial overlap variations, thus leading to extensive matching possibilities. Then, we prune this matching set to guarantee that, for any 2 superpixels respectively belonging to frames t and $t+1$, the number of matching candidates of at least one of them is inferior to 2. Pruning favors the best matches.

C. Spatio-Temporal Tubes grouping

The STT set segmentation consists of 2 passes. The first pass aims to cluster STTs according to color and motion through a meanshift [24] over-segmentation. The second pass iteratively fuses the yielded supertubes, mostly on the basis of their motion similarities. The complete hierarchical fusion tree is stored.

Meanshift segmentation is performed over a (7×number of frames)-D space defined by the 3-color and 4-motion values corresponding to the set of STTs. STTs comparison is performed over their intersecting time.

The second pass takes as input the over-segmented set of supertubes generated by the meanshift segmentation and aims to iteratively group two neighboring supertubes that share the motion similarities. Supertubes as well as the fusion tree structure are stored. For practical purposes, we only keep them from the fusion threshold T till only one supertube remains. This hierarchical grouping preserves the hierarchical ordering of relationships among supertubes.

Hierarchical grouping is achieved by considering all possible fusions. Potential associations are ranked according to their similarities and most similar supertubes are iteratively fused.

V. Volume-aligned Descriptor

To leverage the motion information, local descriptors are densely computed on 3D volume outputted by the hierarchical segmentation. Each video is described as follows. Each supertube is temporally divided in n-frame volumes, ensuring a 50% overlap between consecutive volumes. Any supertube shorter than 5 frames or with motion intensity inferior to 1 pixel per frame is considered unreliable and discarded. For our experiments, we choose $n=7$, in order to produce a large amount of descriptors for richer representation.

Each one of these temporal volumes v belonging to a

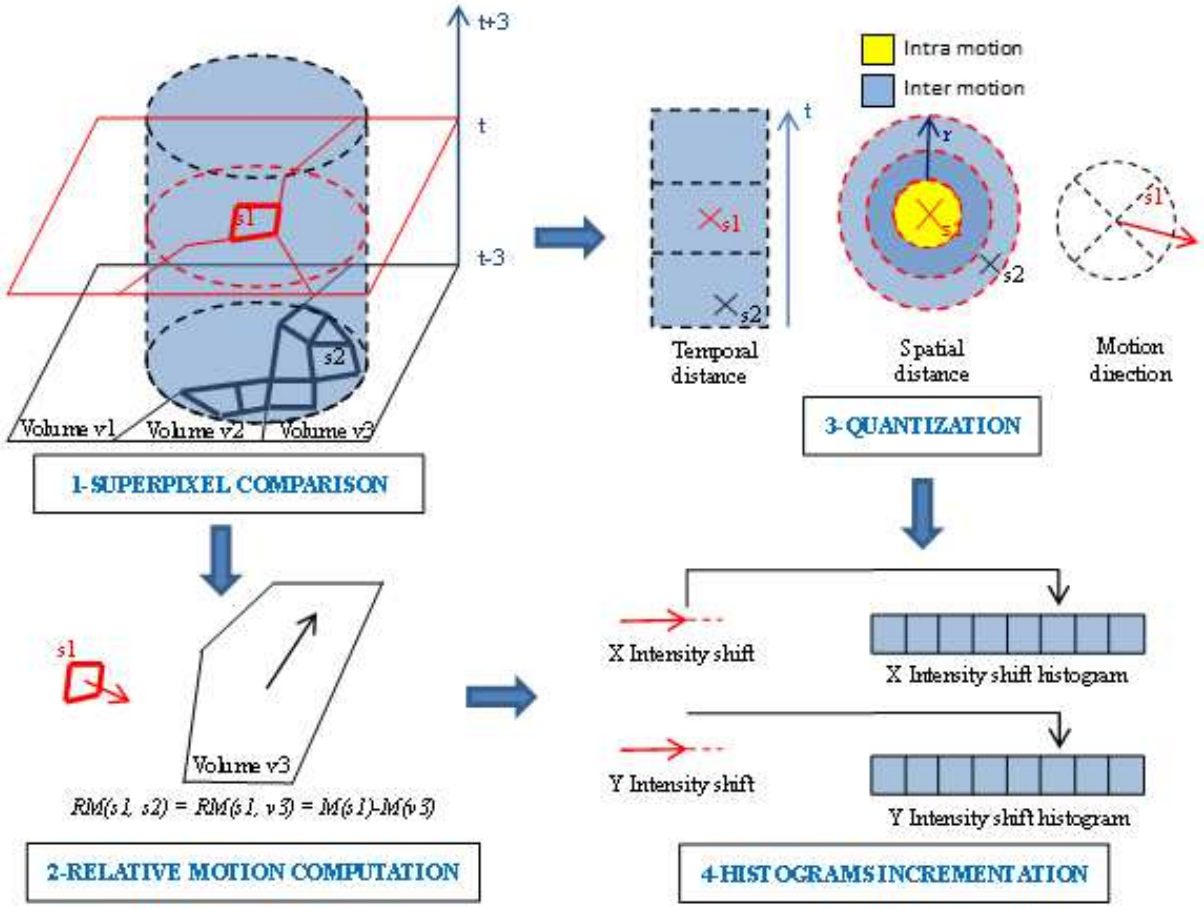

Fig. 3: Relative motion histogram computation. 1-Superpixel comparison comparing the boundary superpixels of the analyzed volume with all the superpixels of other volumes within a defined spatio-temporal neighborhood. 2-relative motion computation associating the neighboring volume motion to its underlying superpixels. 3- Quantization according to temporal distance, spatial distance, and direction. 4- intensity and angle histograms are incremented proportionally to the shifts.

supertube $v(s)$ is then characterized with its underlying superpixel motion and color patterns. For each temporal volume, we built 3 histograms.

First, a 36-values L, a, b color histogram from the CIELAB colorimetric system.

Second, a 72-values modified HoG3D [2] histogram. In our version, the temporal gradient is computed on flow values instead of color ones. Also, to further increase discriminability and reduce computation, only the border superpixels of the analyzed volume v are considered.To speed up the computation, necessary values are pre-processed for each superpixel, and volume HoG3D are then calculated as a combination of included superpixel values. Finally, a 72-values relative motion histogram (RMH) that represents each volume motion differences in terms of intensity with its neighboring supertubes for X and Y axis independently. Contrarily to [13], these relationships are directly inferred at the descriptor level. Calculation is performed according to superpixels from different supertubes lying within a spatial radius r and a temporal radius t. For our experiments, we chose a radius r of 6% of the video width and a temporal radius t of 5 frames with respect to the volumes temporal boundaries. This parameterization allows rich description while limiting the computation to local values. We compare the analyzed volume superpixels motion value to the neighboring supertubes mean motion values. This last approximation increases the robustness to optical flow noise and artifacts and allows volume self-comparison. Finally, to further increase discriminability and reduce computation, only the border superpixels of the analyzed volume are considered. Each pairwise comparison increments the y histogram proportionally to the magnitude of the shift.

The values are quantized according to the analyzed volume motion vector direction, temporal distance, and spatial distance, for a total of respectively (4 $\times 3 \times 3$) 36 relative intensity bins for each X and Y axis. The direction estimation is shifted by 45 degrees to model up, down, left and right directions properly. In order to have a discriminative representation bearing both intra and inter-motion relationships, spatial distance binning is performed

as follows. The first bin encompasses the comparison of *v* with *v(s)* and its sons; whereas the last 2 bins refer to motion differences with other supertubes lying within a distance *r* outside of the volume boundaries. The RMH computation process is illustrated in figure 3.

This Leads to a 180 dimensions descriptor that is further normalized to add up to 1.

The main strength of these descriptors is their adaptability to the scene variations since they are grounded on a video segmentation. This makes them naturally robust to most of the common difficulties: first and foremost the scene topology changes, but also the speed variation, and illumination changes. Moreover, hierarchical supertubes are extracted at different scales or for different parts of the same object depending on its intra-motion variability; it yields a set of descriptors robust to scale changes and occlusions.

The set of descriptors is then quantized according to a 1000 codeword dictionary, commonly employed for activity recognition tasks. Context modeling is performed by describing the video with 9 distinct codeword histograms depending on the spatial size of their descriptor layouts as well as their motion intensity. Sizes usually correspond to the background, object and object parts supertubes. Motion intensity is also a distinctive cue to separate background from objects performing actions of interest. Descriptors stemming from a supertube with an average spatial size / video spatial size ratio of respectively [0, 0.125], [0.125, 0.5], [0.5, 1] will increment their corresponding histograms. The 9 histograms are concatenated, and then only normalized, as the amount of descriptors of each size is a distinctive cue. This yields a rich 9000-value histogram for each video.

VI. Experimental validation

This section presents our test-bed, details the set of experiments and discusses the method's sensitivity to parameters.

A. Database

We use the 15 categories from the TRECVID MED11 challenge set event kit [14] for our experiments. This dataset contains 1941 web-crawled, realistic videos without any negative examples. Challenges include blur, low resolution camera jitter or deformations, brisk zoom, complex character movements, illumination changes, crowds and animals. The 15 event classes are shown in tables I and II.

Evaluated method histograms are computed and fed into a SVM using a χ^2 kernel, and results are evaluated according to the mean average precision metric (mAP) using a 70/30 train/test ratio.

B. Results

We compare our descriptor with state-of-the-art features. We use a different data split than [29] as it isn't available. results are displayed in table I. STT descriptors outperform standard features with a mean average precision of 57.6% and achieve the best category results for 5 out of 15 event categories. Results are best for categories with fast-paced motion patterns such as 'board trick' and lower for textured objects and slow motion patterns like 'sewing project' or 'grooming an animal' categories. Coupling it with a texture or gradient descriptor might lead to further improvement. The mean average computation time, calculated over 50-frames video-clips from all event categories, is 43.8 seconds per clip on a 3.5GHz Intel Xeon, 42.7% of the computation being dedicated to the descriptor extraction. Processing time strongly depends on the scene complexity and the hierarchical tree structure. Up to 1500 frame-clips can be processed with 24Gb of RAM.

C. Parameter Influence Analysis

This subsection gets down to the analysis of our descriptor main constituents, namely the RMH, color, and modified Hog3D histogram combinations, as well as context modeling. Each component's influence on final results is evaluated by running the classification without it and gauging the drop in performance. Results are displayed in table II. The first three columns relate to the 3 descriptor histograms, whereas the last one refers to the context modeling. It shows that, while all 3 histograms are important for classification, modified HoG3D have a greater influence. Moreover, the use of context modeling leads to a 3.5% mAP improvement. This further validates our approach choices.

VII. Conclusion

We have presented a new video descriptor exploiting a hierarchical video segmentation that produces object or object part related volumes at multiple scales. This descriptor embeds motion, appearance, pairwise motion relationships, and context modeling. It is robust and scalable. The approach has been validated on the event recognition task through experimental results on the TRECVID MED11 dataset.

We also provided a new approach that describes a video in a combined statistical and structural manner. In the future, we aim to build on this last characteristic to prune independent, yet meaningful video parts by having a greater understanding of their relationships. Another possible direction to investigate is to perform sequence reasoning on the set of descriptors originating from the same supertube to improve matching between video clips.

Acknowledgment

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC0067. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies

Features	GIST from [29]	STIP from [29]	Dollar from [29]	MFCC from [29]	SIFT from [29]	ISA from [29]	MOSIFT	STT descriptor
Board trick	0.581	0.548	0.543	0.494	0.524	0.653	0.717	**0.763**
Feeding an animal	0.286	0.301	0.275	0.343	0.391	**0.514**	0.359	0.395
Landing a fish	0.462	0.36	0.398	0.291	**0.698**	0.646	0.615	0.679
Wedding	0.521	0.622	0.632	0.534	0.663	0.665	**0.759**	0.699
Woodwork	0.241	0.308	0.319	0.458	0.473	0.559	0.398	**0.769**
Birthday party	0.282	0.202	0.353	**0.632**	0.476	0.59	0.625	0.463
Changing a tire	0.076	0.174	0.191	0.203	0.295	**0.45**	0.366	0.261
Flash mob	0.607	0.729	0.783	0.627	0.808	0.785	**0.838**	0.811
Getting a vehicle unstuck	0.395	0.395	0.482	0.353	0.482	0.441	0.628	**0.763**
Grooming an animal	0.277	0.283	0.288	0.249	0.362	0.497	**0.626**	0.38
Making a sandwich	0.219	0.196	0.19	0.254	**0.325**	0.285	0.305	0.259
Parade	0.367	0.36	0.452	0.423	0.463	0.422	0.599	**0.793**
Parkour	0.341	0.438	0.354	0.271	0.678	0.620	0.658	**0.835**
Repairing an appliance	0.632	0.49	0.579	0.776	0.638	**0.702**	0.604	0.551
sewing project	0.206	0.23	0.327	0.378	0.351	**0.553**	0.429	0.214
MEAN AP	0.3662	0.3757	0.4111	0.4191	0.5085	0.5588	0.5684	**0.576**

TABLE I: Comparative results of our method with various descriptors on the TRECVID MED11 event kit, using mAP metric. The 15 first rows relate to independent event category score. Last row depicts the average score over the 15 categories. Best scores are shown in bold.

Features	STT no RMH	STT no Hog3D	STT no color	STT no context
Board trick	0.697	0.653	0.753	0.651
Feeding an animal	0.322	0.315	0.365	0.311
Landing a fish	0.661	0.635	0.646	0.613
Wedding	0.629	0.686	0.611	0.673
Woodwork	0.678	0.636	0.711	0.426
Birthday party	0.434	0.438	0.458	0.531
Changing a tire	0.259	0.251	0.259	0.234
Flash mob	0.798	0.739	0.802	0.804
Getting a vehicle unstuck	0.667	0.635	0.687	0.608
Grooming an animal	0.359	0.329	0.375	0.265
Making a sandwich	0.235	0.245	0.258	0.336
Parade	0.719	0.654	0.642	0.707
Parkour	0.689	0.679	0.735	0.521
Repairing an appliance	0.508	0.481	0.549	0.628
sewing project	0.211	0.212	0.209	0.358
MEAN AP	0.524	0.506	0.537	0.511

TABLE II: Independent evaluation of STT descriptor components. The 15 first rows relate to independent event category score. Last row depicts the average score over the 15 categories.

or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

REFERENCES

[1] I. Laptev and T. Lindeberg, *Space-time interest points,* ICCV, 2003.

[2] N. Buch, J. Orwell, S.A. Velastin, *3D Extended Histogram of Oriented Gradients (3DHOG) for Classification of Road Users in Urban Scenes,* BMVC, 2009.

[3] M.-Y. Chen and A. Hauptmann, *MoSIFT: Recognizing Human Actions in Surveillance Videos,* CMU-CS-09-161, Carnegie Mellon University, 2009.

[4] H. Bay, A. Ess, T. Tuytelaars, L. van Gool, *Speeded-up Robust Features (SURF),* CVIU, Vol. 110, No. 3, pp. 346-359, 2008.

[5] P. Scovanner, S. Ali, and M. Shah, *A 3-Dimensional SIFT Descriptor and Its Application to Action Recognition,* ACM Multimedia, 2007.

[6] L. Yeffet and L. Wolf, *Local Trinary Patterns for Human Action Recognition,* ICCV, 2009.

[7] Ross Messing, Christopher J. Pal, Henry A. Kautz, *Activity recognition using the velocity histories of tracked keypoints,* ICCV 2009.

[8] B. D. Lucas and T. Kanade, *An iterative image registration technique with an application to stereo vision,* Proceedings of Imaging Understanding Workshop, pages 121-130, 1981.

[9] J. Sun, X. Wu, S Yan, L.-F. Cheong, T.S. Chua, and J. Li, *Hierarchical spatio-temporal context modeling for action recognition,* CVPR, 2009.

[10] H. Wang, A. Klser, C. Schmid, and L. Cheng-Lin, *Action Recognition by Dense Trajectories,* CVPR, 2011.

[11] N. Dalal, B.Triggs and C.Schmid, *Human Detection Using Oriented Histograms of Flow and Appearance,* ECCV, 2006.

[12] M. Raptis, and S. Soatto, *Tracklet Descriptors for Action Modeling and Video Analysis,* ECCV, 2010.

[13] Y.-G. Jiang, Q. Dai, X. Xue, W. Liu, and C.-W. Ngo, *Trajectory-Based Modeling of Human Actions with Motion Reference Points,* ECCV, 2012.

[14] http://www.nist.gov/itl/iad/mig/med11.cfm

[15] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, C. Schmid, *Evaluation of local spatio-temporal features for action recognition,* BMVC 2009.

[16] A. Noguchi and K. Yanai, *A SURF-Based Spatio-Temporal Feature for Feature-Fusion-Based Action Recognition,* ECCV, 2010.

[17] P. Mattikainen, M. Hebert, and R. Sukthankar, *Representing Pairwise Spatial and Temporal Relations for Action Recognition,* ECCV, 2010.

[18] N. Ballas, B. Delezoide, F. J. Prlteux, *Trajectory signature for action recognition in video,* ACM Multimedia, 2012.

[19] D. G. Lowe, *Distinctive image features from scale-invariant keypoints,* International Journal of Computer Vision, 60:91110, 2004.

[20] E. Mortensen, H. Deng, and L. Shapiro, *A SIFT descriptor with global context,* CVPR, 2005.

[21] C. Li and L. Ma, *A new framework for feature descriptor based on SIFT,* Pattern Recogn. Lett., 30(5):544557, 2009.

[22] G. Carmichael, R. Laganire, and P. Bose, *Global Context Descriptors for SURF and MSER Feature Descriptors,* CRV, 2010.

[23] J. Matas, O. Chum, M. Urban, and T. Pajdla, *Robust wide baseline stereo from maximally stable extremal regions,* BMVC, 2002.

[24] D. Comaniciu, P. Meer: Mean Shift: A Robust Approach toward Feature Space Analysis, *Distinctive image features from scale-invariant keypoints,* PAMI, Vol. 24, No. 5, 603-619, 2002.

[25] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, *SLIC Superpixels,* EPFL Technical Report no. 149300, June 2010.

[26] F. Galasso, R. Cipolla, B. Schiele, *Video segmentation with superpixels,* ACCV, 2012.

[27] M. Werlberger, T. Pock, and H. Bischof, *Motion Estimation with Non-Local Total Variation Regularization,* CVPR, 2010.

[28] R.Trichet, R. Nevatia, *Video Segmentation with spatio-temporal tubes,* AVSS 2013.

[29] H. Izadinia, and M. Shah, *Recognizing Complex Events using Large Margin Joint Low-Level Event Model,* ECCV, 2012.

[30] M. Marszaek, I. Laptev, C. Schmid, *Actions in context,* CVPR, 2009.

[31] M. Reso, J. Jachalsky, B. Rosenhahn, J. Ostermann, *Temporally Consistent Superpixels,* ICCV, 2013.

[32] J. Chang, D. Wei, J. W. Fisher III, *A Video Representation Using Temporal Superpixels,* CVPR, 2013.

[33] A. Levinshtein, C. Sminchisescu, S. Dickinson, *Optimal Image and Video Closure by Superpixel Grouping,* IJCV, 2012.