Decreasing the Number of Features for Improving Human Action Classification

Jacques de Souza Kleber, Arnaldo de Albuquerque Araújo, Zenilton Kleber G. Do Parocinio Jr, Jean Cousty, Laurent Najman, Yukiko Kenmochi, Silvio Jamil F. Guimarães

Decreasing the Number of Features for Improving Human Action Classification

Kleber Jacques de Souza*, Arnaldo de Albuquerque Araújo *, Zenilton Kleber G. do Patrocínio Jr.‡,
Jean Cousty†, Laurent Najman†, Yukiko Kenmochi† and Silvio Jamil F. Guimarães‡

*NPDI/DCC/UFMG
Federal University of Minas Gerais - Computer Science Department
Belo Horizonte, MG, Brazil
{arnaldo,kleberjacques} @dcc.ufmg.br
†Université Paris-Est, Laboratoire d'Informatique Gaspard-Monge UMR 8049,
UPEMLV, ESIEE Paris, ENPC, CNRS, F-93162 Noisy-le-Grand France
{j.cousty,y.kenmochi,l.najman} @esiee.fr
‡Audio-Visual Information Proc. Lab. (VIPLAB)
Computer Science Department – ICEI – PUC Minas
{zenilton, sjamil}@pucminas.br

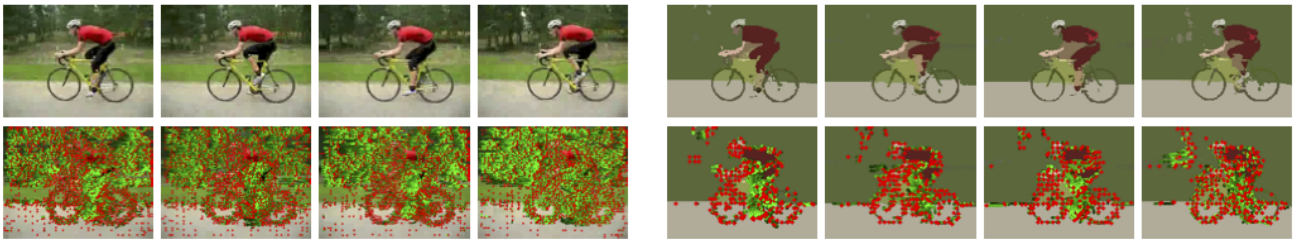

Fig. 1. Example of feature extraction in the UCF YouTube Action dataset using dense trajectories. The results for original frames are illustrated on the left and the ones for segmented frames are on the right.

***Abstract*—Action classification in videos has been a very active field of research over the past years. Human action classification is a research field with application to various areas such as video indexing, surveillance, human-computer interfaces, among others. In this paper, we propose a strategy based on decreasing the number of features in order to improve accuracy in the human action classification task. Thus, to classify human action, we firstly compute a video segmentation for simplifying the visual information, in the following, we use a mid-level representation for representing the feature vectors which are finally classified. Experimental results demonstrate that our approach has improved the quality of human action classification in comparison to the baseline while using 60% less features.**

***Keywords*-Spatio-temporal video segmentation; human action classification; BossaNova representation**

I. Introduction

In this paper, we address the task of Human Action Classification which is the process of naming human actions based on the video content. Therefore, it can be defined as: given a pre-determined number of actions, we need to classify a new action in one of these types. Many works address this problem using two stages [1]: (i) feature extraction; and (ii) action classification. Feature extraction is the main vision task in action classification and consists in extracting visual information from the video. Action classification involves the steps of learning statistical models from the extracted features, and using those models to classify new feature observations.

In a typical approach to human action classification, the stage of feature extraction is performed directly on the raw data (pixels of the video), and it may contain noise or irrelevant information. Therefore, this paper presents an approach to human action classification based on using video segmentation for decreasing the number of features while increasing the accuracy of classification task. The main idea here is to filter out unnecessary information and noise that may tamper with the classification process. Recently, there has been a growing trend of using temporal video segmentation as preprocessing for action recognition [2], [3], [4]. It was hoped that segmentation methods could partition videos into coherent constituent parts, in such a way that recognition could then be simply carried out based on the obtained segments. Niebles et al. [2] proposed a strategy for modeling temporal structure of decomposable motion segments for activity classification. They used a discriminative model that encodes a temporal decomposition of video sequences, and appearance models for each motion segment. In [3], the authors proposed a new representation of local spatio-temporal cuboids based

on atomic actions that represent the basic units of human actions. In [4], the authors presented a motion descriptor for human action recognition that is based on both the accordion representation of the video and its temporal segmentation into elementary motion segments.

Spatial-temporal motion and appearance context information around pixels can deliver more complex motion and appearance structures. In [5], the authors proposed a motion boundary based dense sampling strategy, called DT-MB, to reduce the number of trajectory yet preserve the power of dense trajectory using a group of spatial temporal context descriptor. Likewise, Yi and Lin [6] proposed a mid-level approach to represent and model the spatio-temporal relationship of video elements for the purpose of human activity classification in unconstrained environments. In [7], the authors proposed the use of tubelets, i.e, mid-level representation from successive mergings of the spatio-temporal segmentations, to perform action localization.

The main contributions of this work are twofold: (i) proposal of a method for decreasing the number of features that will be used for video classification; and (ii) improvement of the accuracy in the task of human action classification. It is important to note that these contributions were possible due to the replacement of the cuboids by spatio-temporal segments. And their importance are related to the fact that they help reducing training time during classification step, and also improving the quality of video classification (as it will be shown further ahead), which are responsible for generating classifiers with better performance despite the fact that a smaller number of features was used in this process.

This work is organized as follows. In Section II, we present the approach of human action classification used in this work. Then, experimental results are presented in Section III. Finally, Section IV presents final remarks and discusses possible research lines for future works.

II. Action Human Classification

In human action classification, the common approach is to extract image features from the video and to issue a corresponding action class label. However in this work we address this task according to the method illustrated in Fig. 2.

A. Spatio-Temporal Video Segmentation

The interpretation of video data is a complex activity, so a step of segmentation may be necessary to partition the video into structures with relevant semantic content to aid in the analysis process. There are in the literature several algorithms for video segmentation. Some of these algorithms simply apply techniques for image segmentation to the video frames without considering temporal coherence [8], [9], while others can preserve the temporal information as supervoxels, which is a set of spatially contiguous voxels (a voxel has three coordinates (x, y, t), in which time is represented as the third dimension) that have similar appearance (intensity, color, texture, etc.) [10].

Hierarchical video segmentation provides region-oriented scale-space, i.e., a set of video segmentations at different detail levels in which the segmentations at finer levels are nested with respect to those at coarser levels. Hierarchical methods have the interesting property of preserving spatial and neighboring information among segmented regions. Hierarchical video segmentation generalizes these concepts in order to consider spatiotemporal regions exhibiting in both appearance and motion. The benchmark and library LIBSVX proposed in [11] contains several *state-of-the-art* methods for early hierarchical video segmentation. The implementations of the methods GB [12], GBH [13], MeanShift [14] and SWA [15] applied to video segmentation are available in LIBSVX.

In [16], HOScale method is used to obtain good results for video segmentation and it easily computes any hierarchical segmentation level. The authors in [16] propose a video segmentation method that is not dependent on the hierarchical level; and, consequently, it is possible to compute any level without computing the previous ones. Thanks for that, the time for computing a segmentation is almost the same for any specified level since the video segmentation problem is transformed into a graph partitioning problem in which each part will correspond to one region of the video.

In this work, we use the HOScale method [16] to perform spatio-temporal video segmentation as preprocessing for action classification, thus we can easily define hierarchical level to be used in the action classification process, since it is not necessary to compute all lower (finer) segmentations. Moreover, according to experimental results presented in [16], HOScale produces good quantitative and qualitative results when compared to other methods.

In Fig. 3, we present examples of video segmentations for a video belonging to UCF YouTube Action dataset [17], in which it is possible to verify the change in details of visual content represented by different hierarchical levels. The color of each supervoxel is the average of the colors regarding the original values of the voxels.

As can be seen in Fig. 3, depending on the selected hierarchical level, we can control the quantity of details of the scene represented in each image, which can be useful to

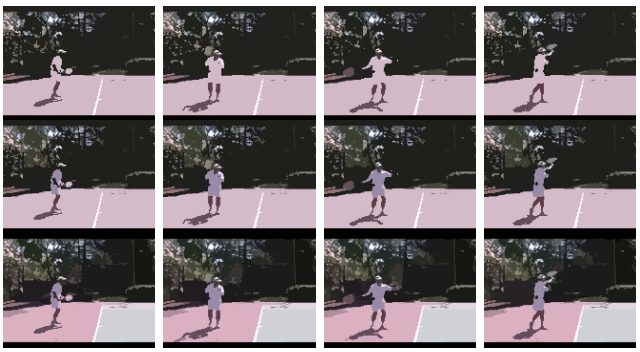

Fig. 3. Examples of video segmentations for a video belonging to UCF YouTube Action dataset [17]. The rows (from top to bottom) illustrate the results obtained by HOScale [16] with different hierarchical levels.

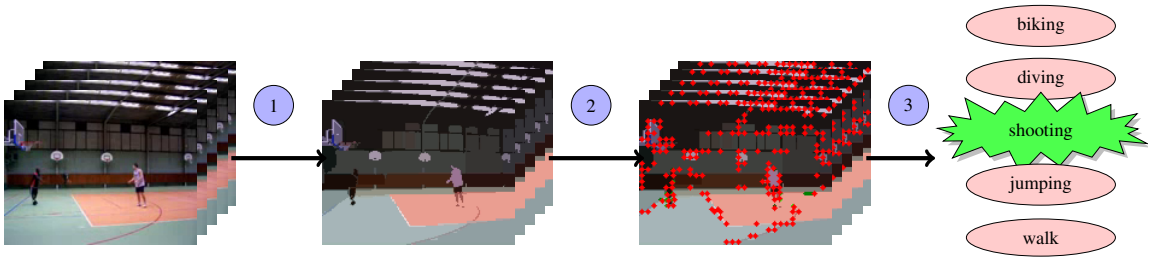

Fig. 2. Outline of our method: the video is segmented (step 1); the features are extracted from the segmented video (step 2); and finally, the identification of human action classification is made (step 3).

remove noise and redudant information.

B. Feature Extraction

Aiming at action classification, different representations can be used for extracting visual information from the video like HOG (Histogram of Oriented Gradients)/HOF (Histogram of Optical Flow) [18], Cuboids [19] and Extended SURF [20].

Motion descriptors are well suited to describe human actions [21]. HOF descriptors characterize local motions. They are computed by dividing the space time neigborhood of the Harris3D interest points into small space-time regions and accumulating a local 1-D histogram of optic flow over the pixels of each sub-region. Dalal et al. [22] proposed the motion boundary histograms (MBH) descriptor for human action detection. The MBH descriptor describes the relative motion between pixels by computing the gradient of the optical flow. In [21], MBH is used as motion descriptor for dense trajectories.

We have no intentions to propose changes in existing feature extraction approaches, but only to apply them to the segmented video data. After the feature extraction step, the Bag-of-Words (BOW) model is used to organize low-level features to represent each video. This approach commonly consists of two phases, i.e., feature coding and feature pooling. In this work, we use BossaNova, a representation for content-based concept detection in images and videos, which enriches the Bag-of-Words model [23]. The BossaNova approach follows the Bag-of-Words (BoW) formalism (coding/pooling), but proposes a video representation which keeps more information than BoW during the pooling step. Thus, the BossaNova pooling estimates the distribution of the descriptors around each codeword, by computing a histogram of distances between the descriptors found in the video and those in the codebook. More details can be found in [23].

In [23], the authors applied their representation to image recognition. In comparison to the BoW, BossaNova significantly outperforms it. Furthermore, by using a simple histogram of distances to capture the relevant information, the method remains very flexible and keeps the representation compact. For those reasons, we chose the BossaNova approach as the mid-level feature to be used in the experiments.

III. Experimental Results

In order to evaluate the proposed approach, we performed some human action classification experiments by using two well-known datasets: KTH and UCF YouTube Action.

A. Datasets and Experimental Setup

The KTH dataset [24] consists of six human action classes. The background is homogeneous and static in most sequences. In total, the data consists of 2,391 video samples. We followed the original experimental setup of the authors in [24] in the classification step, we trained and evaluated a multiclass classifier and reported average accuracy over all classes.

The UCF YouTube Action dataset [17] contains 11 action categories. This dataset is challenging due to large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background and illumination conditions. The dataset contains a total of 1,168 sequences. We followed the original setup [17], using leave-one-out cross-validation for a predefined set of 25 groups. Average accuracy over all classes is reported as the performance measure.

Regarding the video segmentation, we used HOScale method [16] because it presented good results for video segmentation and it easily computes any hierarchical segmentation level. Moreover, according to experimental results presented in [16], HOScale produced good quantitative and qualitative results when compared to other *state-of-the-art* approaches, such as, GB [12], GBH [13], MeanShift [14] and SWA [15]. That is the main reason behind our decision of using this method in our strategy for human action classification. Because of the simplicity of the scene, we chose the low number of supervoxels to segment the video. In order to simplify the information to be processed, since the intention is to classify human action, the focus should be on the shape and movement of people in the scene.

Regarding the feature descriptor, we have chosen to use an approach with a dense descriptor (dense trajectories [21]) because it is simple and achieved good results. After the feature extraction step, the Bag-of-Words (BOW) model is used to organize the low-level features to represent each video using the mid-level BossaNova representation. Here, we used the following BossaNova parameter values: $B = 2$, $\lambda_{min} = 0.4$, $\lambda_{max} = 2$, $s = 10^{-3}$ and $M = 4000$ (number of visual codewords).

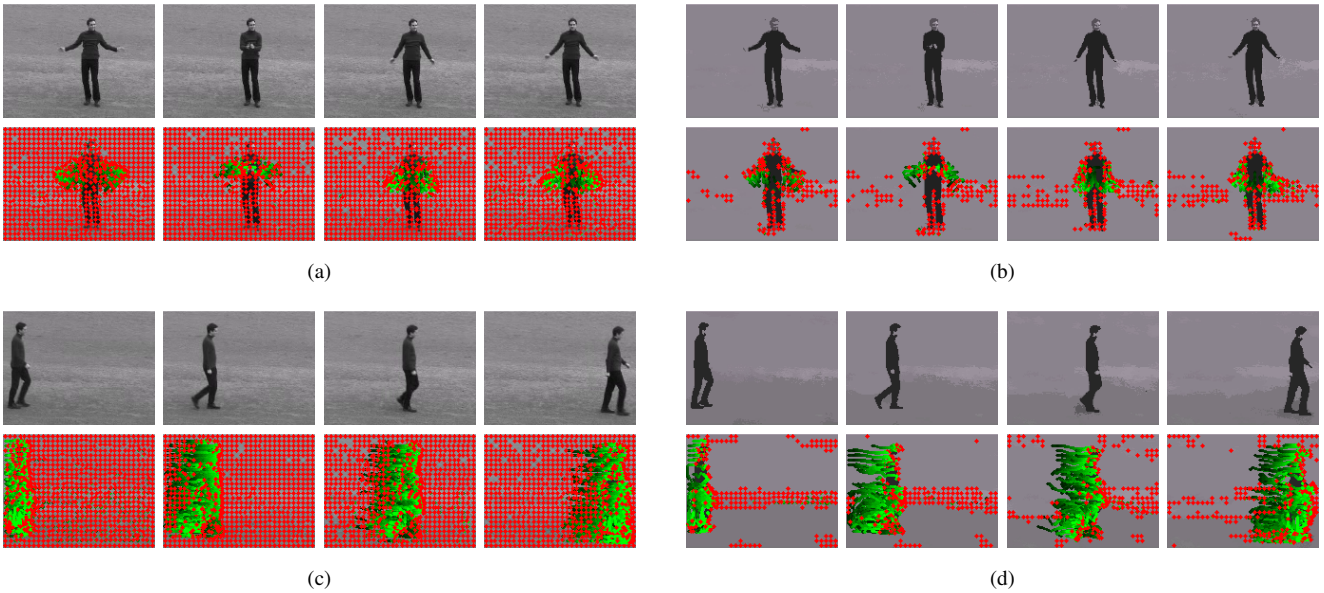

Fig. 4. Example of feature extraction in the KTH dataset using dense trajectories [21]. The results for original frames are illustrated in (a) and (c), while the results for segmented frames is presented in (b) and (d).

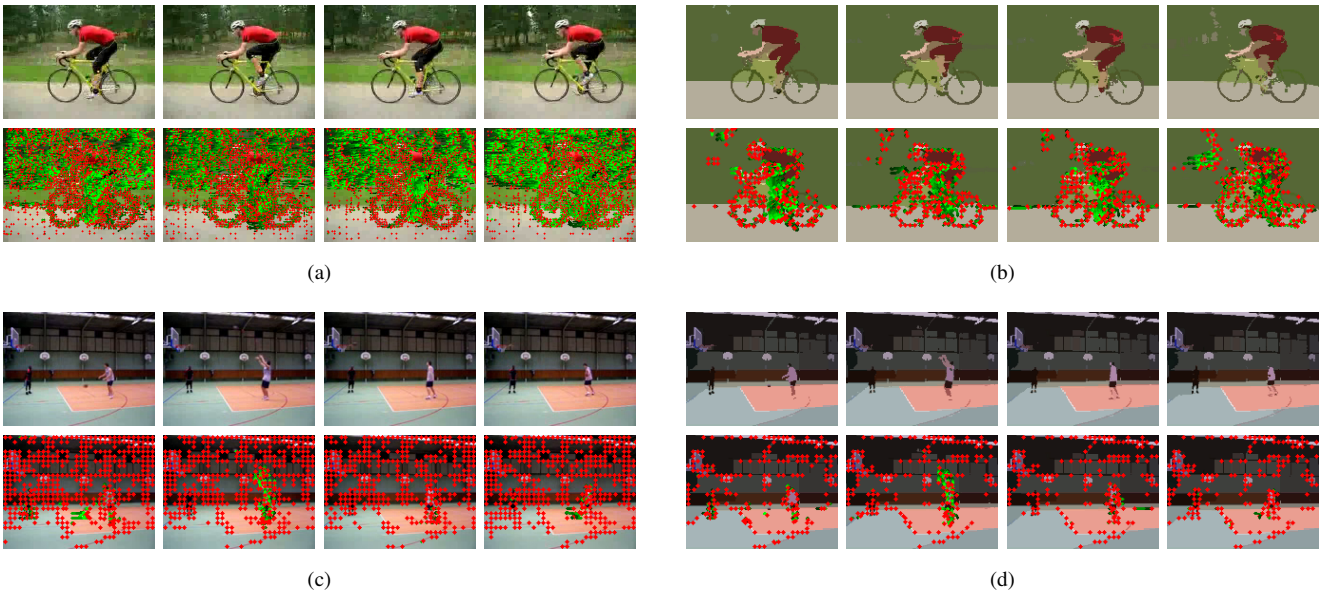

Fig. 5. Example of feature extraction in the UCF YouTube Action dataset using dense trajectories [21]. The results for original frames are illustrated in (a) and (c), while the results for segmented frames is presented in (b) and (d).

For classification we use non-linear SVM with an RBF-$\mathscr{X}^2$ kernel which is a popular classifier that is used throughout different works for humam action classification [21], [18]. This classifier is the most used in human action classification and it is therefore interesting to make fair comparisons between different approaches.

B. Visual Analysis

To extract the features we used the dense trajectory and motion boundary descriptors [21] (examples can be seen in Fig. 4 and 5) which use dense detector approach. The algorithms extract various features, i.e., trajectory, HOG, HOF and MBH, and the overall combination using the multi-channel approach. The results for MBH are obtained by combining MBHx and MBHy channels.

One can easily see the reduction of the number of features extracted from segmented videos (see Fig. 4(b),Fig. 4(d), Fig. 5(b) and Fig. 5(d)). This has an impact on reducing training time, since less data is used. Moreover, it has also contributed to improve the quality of data used.

C. Quantitative Assessment

As mentioned before, to evaluate our approach, we used two different Datasets: KTH and UCF YouTube Action. In Tables I and II, we summarize some results. It is easy to see that our approach presented good results when compared to some *state-of-the-art* methods for the human action classification task.

In Fig. 6, we present two charts. The first one is related to the number of features which are used in the classification task and the second one concerns to the accuracy of our method. In both cases, we have studied the behaviour of our strategy concerning several number of computed supervoxels.Because of the simplicity of the scene, we choose the low number of supervoxels to segment the video. In order to simplify the information to be processed, since the intention is to classify action human the focus should be on the form and movement of people in the scene. For this we performed experiments varying the number of supervoxels between 25 and 100. The best result was using 75 supervoxels, i.e., each video has been simplified to 75 segments.

As we can see in Table I, for KTH dataset with dense trajectory descriptor, the baseline [21] (without video segmentation) presents an accuracy of 94.2%, using $4,649,455$ features (see Table II). The best accuracy obtained by our method is 95.4%, using almost 35% (3,053,780 features, see Table II) less features than the original method. From this experiment, it is important to note that the proposed method improves the accuracy while the number of features is decreased. Table III presents the confusion matrix describing in details the performance of our method for KTH dataset.

Also in Table I, concerning now UCF YouTube Action dataset, the dense trajectory descriptor without video segmentation, can observe that it is possible to correctly classify 84.1%, using 14,535,963 features (see Table II). Otherwise, our method, using spatio-temporal video segmentation it was possible to correctly classify 87.4%, using 5,708,152 features, see Table II (which represents a reduction of 60%).

Through the experiments performed, we can observe that it is possible to improve the human action classification by using video segmentation. This has occurred because video segmentation has grouped pixels into structures with relevant semantic content which aids in the classification process using less information. Actually, the use of video segmentation allows the extraction of a smaller number of features and helps not only reducing training time during classification step, but also improving the quality of video data used, since it filters out unnecessary information and noise. One may argue that those results are due to the capacity of our method in identifying better the support vectors needed to produce higher values of accuracy in human action classification. But, unfortunately, we do not have any real evidence to uphold that claim yet.

IV. Conclusion and Further Work

In the content-based visual information retrieval it is very important to filter the data to be processed, since the presence of noise or irrelevant information may hinder the results. In this work, we proposed a human action classification approach that uses the spatio-temporal video segmentation as a preprocessing step in order to improve the classification process.

The HOScale method [16] is used to perform spatio-temporal video segmentation as preprocessing for action classification; and, to extract the features, we used the dense trajectory and motion boundary descriptors [21]. The Bag-of-Words (BOW) was adopted to organize low-level features to represent each video with the mid-level representation BossaNova [23]. Moreover, for the classification task, we used a non-linear SVM.

Experimental results demonstrated that our approach has improved the quality of human action classification despite the fact that if it has used less features. The experiments were realized on two datasets: KTH and UCF YouTube Action. In KTH dataset using 35% less features, we obtained a little better result than the method without segmentation. Also, in UCF YouTube Action dataset, which is a more complex dataset, using 60% less features, we observed a good increase in the accuracy results.

Furthermore, we used a recent mid-level representation, called BossaNova, which enriches the Bag-of-Words model, to describe the video features, which also showed to be a good approach for video representation as it has been for image representation.

TABLE I
Comparison of accuracy values for tested approaches.

Approach	KTH	UCF YouTube
Wang et al. (2013)[21]	94.2	84.1
Peng et al. (2013)[5]	95.6	86.56
Yi and Lin (2015)[6]	-	84.63
Our method	95.4	87.4

TABLE II
Comparison of the number of features.

Approach	Dataset	# features
Wang et al. (2013)[21]	KTH	4,649,455
	UCF YouTube	14,535,963
Our method	KTH	3,053,780
	UCF YouTube	5,708,152

TABLE III
Confusion matrix describing the performance of our method for KTH dataset.

	boxing	hclapping	hwaving	jogging	running	walking
boxing	97.2%	0	0	0	0	2.8%
hclapping	1.4%	98.6%	0	0	0	0
hwaving	0	5.5%	94.5%	0	0	0
jogging	0	0	0	94.4%	4.9%	0.7%
running	0	0	0	12.5%	87.5%	0
walking	0	0	0	0	0	100%

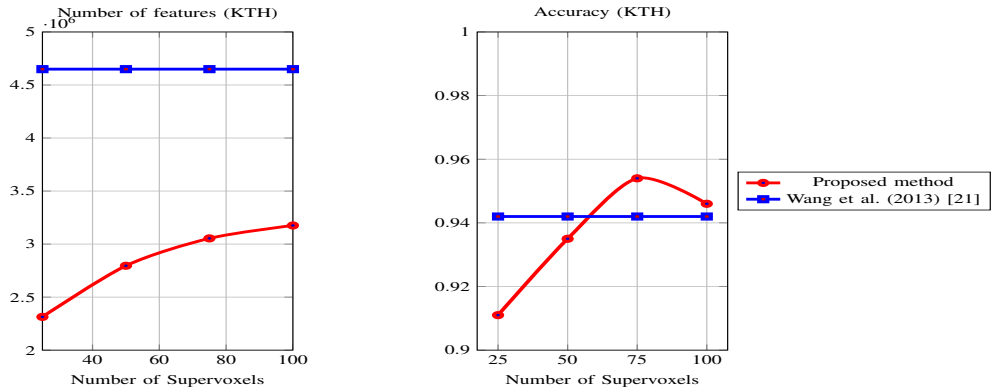

Fig. 6. A comparison between the proposed method and Wang et al. [21] concerning the number of features which are used for human action classification and the accuracy obtained by both methods.

For further works, we will study different ways for extracting features from video segments and we also plan to apply our approach to other databases (and even in other scenarios). Another interesting research path is to investigate the quality of video data used during (and filter out before) training time for the classification step and its relationship with the support vectors needed to produce better accuracy results in human action classification.

ACKNOWLEDGMENT

The authors are grateful to CAPES/COFECUB (592/08), CAPES/PVE (88881.064965/2014-01), FAPEMIG (PPM-006-16 and PPM-177-14), CNPq and PUC Minas for the financial support to this work.

REFERENCES

[1] R. Poppe, "A survey on vision-based human action recognition," *Image and Vision Computing*, vol. 28, no. 6, pp. 976 – 990, 2010.

[2] J. C. Niebles, C.-W. Chen, and L. Fei-Fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *Proceedings of the 11th European Conference on Computer Vision: Part II*, Berlin, Heidelberg, 2010, pp. 392–405.

[3] Q. Zhou and G. Wang, "Atomic action features: A new feature for action recognition," in *Computer Vision - ECCV 2012 Workshops and Demonstrations*, ser. Lecture Notes in Computer Science, 2012, vol. 7583, pp. 291–300.

[4] M. Sekma, M. Mejdoub, and C. Ben Amar, "Human action recognition using temporal segmentation and accordion representation," in *Computer Analysis of Images and Patterns*, ser. Lecture Notes in Computer Science, R. Wilson, E. Hancock, A. Bors, and W. Smith, Eds., 2013, vol. 8048, pp. 563–570.

[5] X. Peng, Y. Qiao, Q. Peng, and X. Qi, "Exploring motion boundary based sampling and spatial-temporal context descriptors for action recognition," in *The 24th British Machine Vision Conference (BMVC)*, 2013, pp. 1–11.

[6] Y. Yi and M. Lin, "Human action recognition with graph-based multiple-instance learning," *Pattern Recognition*, 2015.

[7] M. Jain, J. van Gemert, H. Jegou, P. Bouthemy, and C. Snoek, "Action localization with tubelets from motion," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, June 2014, pp. 740–747.

[8] J. Chen, S. Paris, and F. Durand, "Real-time edge-aware image processing with the bilateral grid," in *ACM SIGGRAPH 2007 papers*, New York, NY, USA, 2007.

[9] H. Winnemoller, S. C. Olsen, and B. Gooch, "Real-time video abstraction," *ACM Trans. Graph*, vol. 25, 2006.

[10] C. Xu, C. Xiong, and J. J. Corso, "Streaming Hierarchical Video Segmentation," in *Proceedings of European Conference on Computer Vision*, 2012, pp. 626–639.

[11] C. Xu and J. J. Corso, "Evaluation of super-voxel methods for early video processing," *Vision and Pattern Recognition (CVPR), 2012*, pp. 1202–1209, Jun. 2012.

[12] P. F. Felzenszwalb and D. P. Huttenlocher, "Efficient Graph-Based Image Segmentation," *IJCV*, vol. 59, no. 2, pp. 167–181, 2004.

[13] M. Grundmann, V. Kwatra, M. Han, and I. Essa, "Efficient Hierarchical Graph Based Video Segmentation," *Ieee Cvpr*, 2010.

[14] S. Paris and F. Durand, "A Topological Approach to Hierarchical Segmentation using Mean Shift," in *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, 2007, pp. 1–8.

[15] E. Sharon, A. Brandt, and R. Basri, "Fast multiscale image segmentation," in *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, vol. 1, 2000, pp. 70–77 vol.1.

[16] K. J. F. de Souza, A. d. A. Araújo, Z. K. G. do Patrocínio Jr., and S. J. F. Guimarães, "Graph-based hierarchical video segmentation based on a simple dissimilarity measure," *Pattern Recognition Letters*, vol. 47, pp. 85–92, 2014.

[17] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos in the wild," *IEEE International Conference on Computer Vision and Pattern Recognition*, 2009.

[18] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conference on Computer Vision and Pattern Recognition, 2008.*, June 2008, pp. 1–8.

[19] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proceedings of the 14th International Conference on Computer Communications and Networks*, ser. ICCCN '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 65–72.

[20] 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*, Berlin, Heidelberg, 2008, pp. 650–663.

[21] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, May 2013.

[22] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Proceedings of the 9th European Conference on Computer Vision: Part II*, Berlin, Heidelberg, 2006, pp. 428–441.

[23] S. Avila, N. Thome, M. Cord, E. Valle, and A. d. A. AraúJo, "Pooling in image representation: The visual codeword point of view," *Computer Vision and Image Understanding*, vol. 117, no. 5, pp. 453–465, 2013.

[24] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3*, ser. ICPR '04. Washington, DC, USA: IEEE Computer Society, 2004, pp. 32–36.