Spatio-Temporal Frames in a Bag-of-visual-features Approach for Human Actions Recognition

Ana Paula B. Lopes[1,2], Rodrigo S. Oliveira[1], Jussara M. de Almeida[1], Arnaldo de A. Araújo[1]

[1]Computer Science Department
Federal University of Minas Gerais – UFMG – Belo Horizonte, MG, Brazil

[2]Exact and Technological Sciences Department
State University of Santa Cruz – UESC – Ilhéus, BA, Brazil

{paula, rsilva, jussara, arnaldo}@dcc.ufmg.br

***Abstract*—The recognition of human actions from videos has several interesting and important applications, and a vast amount of different approaches has been proposed for this task in different settings. Such approaches can be broadly categorized in model-based and model-free. Typically, model-based approaches work only in very constrained settings, and because of that, a number of model-free approaches appeared in the last years. Among them, those based in bag-of-visual-features (BoVF) have been proving to be the most consistently successful, being used by several independent authors.**

For videos to be represented by BoVFs, though, an important issue that arises is how to represent dynamic information. Most existing proposals consider the video as a spatio-temporal volume and then describe "volumetric patches" around 3D interest points. In this work, we propose to build a BoVF representation for videos by collecting 2D interest points directly. The basic idea is to gather such points not only from the traditional frames (xy planes), but also from those planes along the time axis, which we call the spatio-temporal frames. Our assumption is that such features are able to capture dynamic information from the videos, and are therefore well-suited to recognize human actions from them, without the need of 3D extensions for the descriptors. In our experiments, this approach achieved state-of-the-art recognition rates on a well-known human actions database, even when compared to more sophisticated schemes.

***Index Terms*—Human Actions; Bag-of-Visual-Features; Video classification;**

I. Introduction

The recognition of human actions from videos has several interesting and important applications, like improving video content-based indexing and retrieval [1], helping in the identification of abnormal behavior in surveillance environments[2], enhancing human-computer interaction [3], remotely monitoring elderly people [4] or analyzing motion patterns of people with motor problems[5].

There is a vast amount of different proposals to detect and/or recognize human actions in the literature and they can be categorized as shown in Figure 1. The techniques in the left branch of this picture – the model-based approaches – rely on modeling the moving objects in the scene and then trying to associate different model parameter sets to different actions. Model-based solutions can be grouped into three major approaches. The first class of approaches consists of those ones based on explicit models of moving objects, like stick models for the human body or parts-based models.

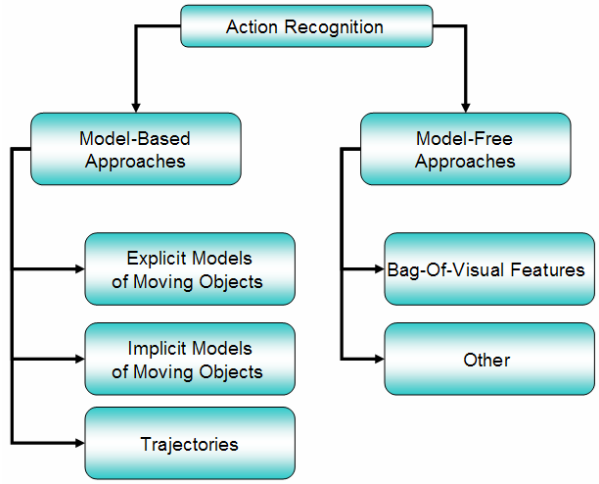


Fig. 1: Categorization of different approaches for human actions recognition.

The second type of model-based approaches for human actions recognition is based on implicit models of the moving objects. In this kind of approach, regions in which the moving object are supposed to be are detected in form of silhouettes or bounding boxes. Then, those selected regions are described in terms of some kind of low level features.

The third category of model-based approaches are those techniques that do not rely on modeling the internal structure of the moving objects, but on the modeling of their trajectories instead. Moving objects can be either the entire human body, body parts or other objects related to the application domain (airplanes, automobiles or even unlabeled moving regions). Recent surveys on implicit and explicit model-based approaches and on trajectory-based ones can be found at [6], [7], [2], [8].

The main drawback of model-based approaches like those just described is their dependance on intermediate tasks to which the available techniques are not reliable enough in

generic situations, like segmentation and tracking. The lack of general solutions to these tasks lead to approaches for human actions recognition that make too many assumptions about the scene, and therefore, are applicable only to very constrained settings. As an example, for spatio-temporal volumetric approaches like the one described in [9], the adequate extraction of silhouettes demands a scene in which the entire body of a unique person against a static and uncluttered background is guaranteed. Of course, so many constraints severely limit the applicability of such a method.

To avoid such limitations, a number of authors has been proposing model-free techniques, in which no previous assumption about the scene content is done. These techniques are represented in the right branch of Figure 1. Most of them are based on some kind of statistical analysis of low level features. Among model-free approaches, those based on bag-of-visual-features (BoVF) have been proved to be the most consistently successful, in a sense that several independent authors have applied BoVF-based techniques with promising results (it is interesting to notice that, due to a lack of standard terminology, BoFV-based approaches have also been denominated bag-of-visual-words, bag-of-keypoints, bag-of-features or bag-of-words in the literature).

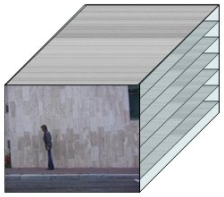
Bag-of-visual-features (BoVF) representations are inspired in traditional bag-of-words (BOW) approaches from textual Information Retrieval. In a BOW, the feature vectors that represent each text document are histograms of word occurrences [10]. Actually, in practice, the words in a BOW are clustered together into families by their roots and only the most discriminative words are taken into account.

The equivalent to those word families in the visual case are small patches clustered by similarity of the descriptors extracted for the patches. Typically, only patches around a sparse selection of interest points are considered. Representations based on a vocabulary of such patches have been used in object recognition, showing good robustness to scale, lighting, pose and occlusion [1].

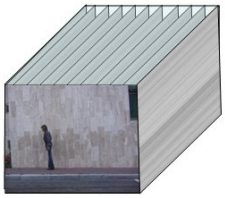
Because of the success of BoVF approaches for object recognition, several authors have been proposed extensions of BoVF representations for videos, most of them aimed at human actions recognition. However, for videos to be represented by BoVFs, an important issue that arises is how to represent dynamic information. Most existing proposals consider the video as a spatio-temporal volume and then describe "volumetric patches" around 3D interest points.

In this work, we propose to build a BoVF representation for videos by collecting 2D interest points directly. The basic idea is to gather such points not only from the traditional frames (xy planes), but also to those planes along the time axis (the spatio-temporal frames). Our assumption is that such features are able to capture dynamic information from the videos, and are therefore well-suited to recognize human actions from them, without the need of 3D extensions for the descriptors. This idea is illustrated in Figure 2 and detailed in Section III. Indeed, as it will be detailed in Section IV, this approach achieved state-of-the-art recognition rates on the well-known Weizmann human actions database [11].

(a) xy planes (traditional frames)



(b) xt planes or "frames"



(c) yt planes or "frames"

Fig. 2: The video as a spatio-temporal volume that can be "sliced" in frames taken along different directions.

This paper is organized as follows: in Section II, some related work is described; in Section III, the proposed approach is presented in detail; in Section IV experimental results using both the Speed-Up Robust Features (SURF)[12] and Scale-Invariant Feature Transform (SIFT) [13] algorithms to select and describe interest points are reported and discussed; finally, concluding remarks are presented in Section V.

II. Related Work

Several authors tried to extend BoFV for human actions classification in videos, in most cases by using spatio-temporal descriptors. In [14], interest points are selected with the Spatio-Temporal Interest Points (STIP) algorithm [15]. The selected points are then described by spatio-temporal jets.

The interest points selection of [16] is based on separable linear filters. In that work, cuboids are defined around the interest points, and several ways of describing these cuboids are tested: *normalized pixel values*, *brightness gradients* and a *windowed optical flow*. In their experiments, brightness gradients provided the best results. Their features – sometimes called Dollar's features – are also used in proposals of [17], [18], [19].

Another BoVF approach for human actions recognition is proposed by [20], which is based on an extension of the SIFT descriptor. The new descriptor adds temporal information extending the original SIFT descriptor to a 3D spatio-temporal space.

In [21], the local descriptors are based on the responses to a bank of 3D Gabor Filters, followed by a MAX-like operation. The BoVF histograms are generated by the quantization of the orientations in nine directions. Instead of a sparse selection of interest points, the features of that work are computed on patches delimited by a sliding window.

In [22], another BoVF representation built from STIP points is proposed. The descriptors are built on the spatio-temporal

volumes around the interest points, by computing coarse histograms of oriented gradients (HoG) and optical flow (HoF).

In [18], a simple BoVF representation based on the brightness gradients features of [16] is used. The main contribution of this work is the application of generative models instead of discriminative ones for classification.

Also using features similar to the ones from [16], [23] proposed the use of a Maximization of Mutual Information (MMI) criteria to merge the cuboid clusters output by k-means. These new clusters are then called Video Words Clusters (VWC).

All these proposals just described build their BoVF representations based on local descriptors which are spatio-temporal extensions of 2D descriptors. The idea behind extending 2D descriptors to 3D spatio-temporal equivalents is to capture local motion information, since because of the intrinsic dynamic aspect of the concepts being classified, motion is an essential clue to differentiate among different human actions.

In this work, based on the observation that the dynamic information is contained in the temporal axis, we reason that it is possible to include the dynamic aspect of the videos by gathering 2D descriptors from the planes formed by one spatial dimension and the temporal one, avoiding the need to create sophisticated 3D extensions to those descriptors. This idea is illustrated in Figure 2 and detailed in Section III.

III. Capturing dynamics from the spatio-temporal frames

A. Building a BoVF representation for videos

The general steps for building a BoVF representation for visual data is depicted in Figure 3. The process starts by selecting a set of points from the videos (step a). This selection can be done densely, by means of a grid applied to the frames or to the spatio-temporal volume. More typical settings, though, apply interest point detectors for a sparser selection, which is going to make the following steps less computationally expensive. The next step (step b) is the description of the region around the selected points. Again, there are a number of alternatives for this step, going from raw gray level values to those more sophisticated descriptors generally delivered by interest point detectors.

Typical descriptors normally have a large dimension and can therefore be submitted to some dimensionality reduction technique, the most common being Principal Component Analysis (PCA) [24]. Dimensionality reduction does not appear in Figure 3 because, although quite common, this is not a mandatory step to create a BoVF.

After points selection and description (with or without dimensionality reduction), the feature space of these points is quantized to form the visual vocabulary (step c). Once the vocabulary is defined, every descriptor on the video is associated with one word from this vocabulary (step d) and then the occurrence for every visual word are counted to form the histogram that constitutes the BoVF representation for that video (step e).

Our BoVF implementation is able to use any point selector/descriptor as input. The vocabulary is created by the k-means clustering algorithm [25], and the vocabulary size k is defined empirically. The final BoVF histogram is normalized to one by computing the relative frequencies of each word.

Given the BoVF representations for the videos, it is possible to apply any classifier for actions recognition. In the present work, a linear SVM classifier (Support Vector Machines) [26] is applied.

B. Collecting Dynamic Information from the Spatio-Temporal Frames

The simplest way to create BoVF representations for videos is by collecting points from every frame in the video segment under consideration and count all them to create a unique histogram for every video segment. The drawback of such a simplistic approach is that it disregards what is happening along the temporal axis, where the dynamic information actually is. As discussed in Section II, some authors have proposed different schemes to include motion information, mostly by extending 2D descriptors to the 3D spatio-temporal space.

In this work, we propose to build a BoVF representation for videos using 2D interest point detector/descriptor algorithms, applied not only to the traditional spatial frames, but also to those planes that we call spatio-temporal frames.

The pure spatial planes (those in the xy direction) are the video original frames. Stacked together, they form a spatio-temporal volume that can be spanned either in x or y directions, forming the "spatio-temporal frames". In other words, the spatio-temporal frames are those planes formed by the temporal axis and one of the spatial axis (xt and yt). This idea is illustrated in Figure 2.

The basic assumption of this proposal is that 2D descriptors extracted from spatio-temporal frames are able to capture the dynamic information contained in the videos, since the evolution over time is now taken into account. The main advantage of this approach is its ability for drawing on existing 2D techniques for feature selection, avoiding the need for 3D extensions for the descriptors.

The BoVF implementation used in this work is kept quite simple in order to focus on the evaluation of this basic idea. Dimensionality reduction is performed by PCA and quantization is done with k-means. No further improvements on the basic BoVF scheme are added and action classification is performed by a linear SVM, with its penalty error parameter fine-tuned each vocabulary size.

IV. Experimental Results

In order to evaluate the ability to capture dynamic information from the spatio-temporal frames as described in the previous section, a set of experiments were performed on the well-known Weizmann human actions database [11]. The Weizmann database is comprised of short video segments, containing nine different people performing ten different actions. The actions considered are *bending to the floor, jumping-jacking, jumping forward, jumping in place, running, walking*

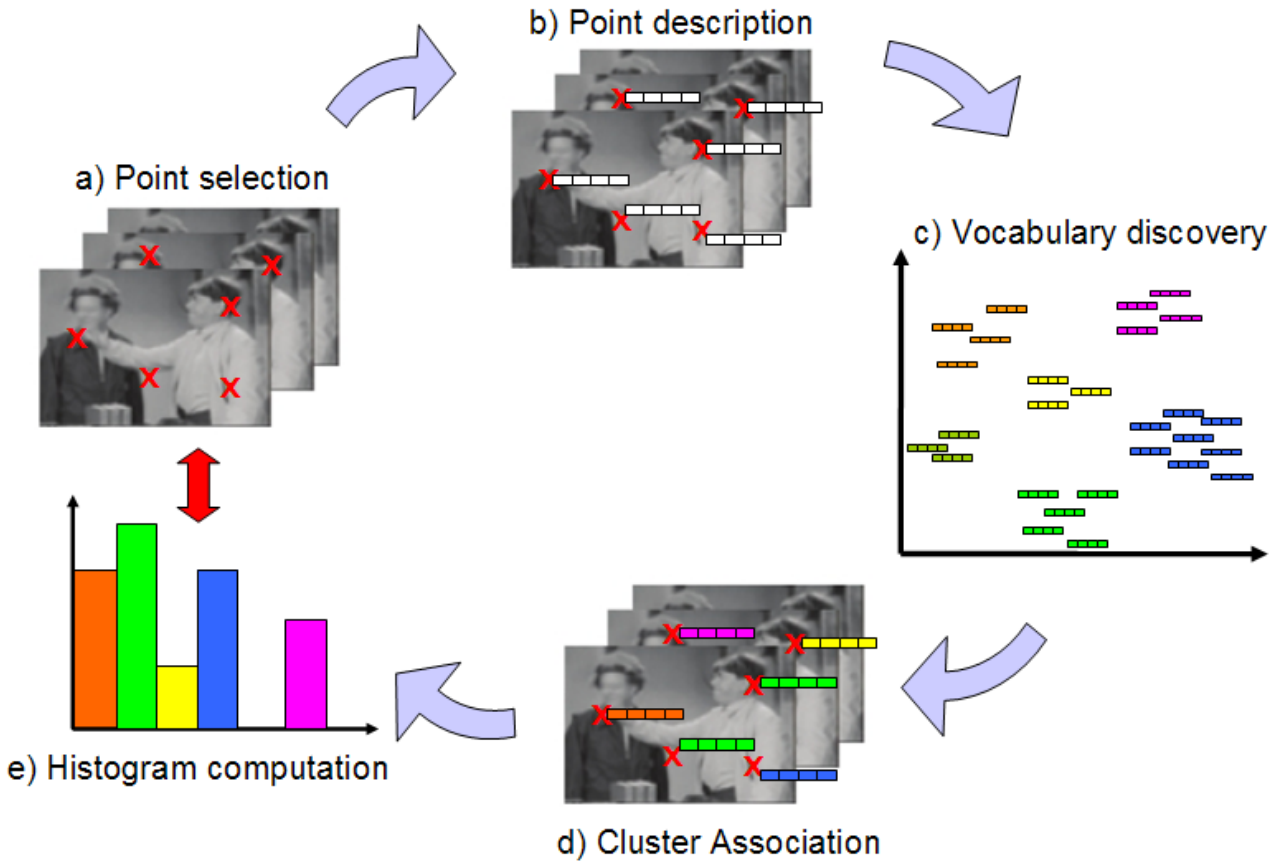


Fig. 3: Steps for creating a Bag of Visual Features for videos. The details of every step are in the text. (This picture is best seen in color).

laterally, jumping on one foot, walking, waving with one hand and with two hands. Snapshots of all actions performed by different people can be seen in Figure 4.

Both the SURF [12] and the SIFT [13] interest point detectors/descriptors are tested. SIFT is an interest point detector and descriptor which looks for points which present invariance to position, scale and location, besides robustness to affine transformations as well as illumination changes. These characteristics turned SIFT interest points quite successful for several Computer Vision tasks, which make this algorithm a natural candidate to provide the points which are going to be the basis of a BoVF representation for images or videos.

SURF algorithm pursues similar goals, but with some simplifications for better performance. SURF's authors claim that it achieves results comparable to SIFT's at a lower computational cost. Since the computational effort for processing videos is always potentially huge, this work proposes to experimentally verify if the claimed similar SURF results still hold in this specific setting.

A. Experimental Setup

Initially, the interest point algorithm is applied to all the frames along each direction. Then, BoVF descriptors are built for the videos using varied combinations of these frames sets[1]. Descriptors extracted from the original (xy) frames form the baseline representation. All possible combinations of frames sets are evaluated for recognition.

For every set of frames and interest points algorithm, the experiments are carried out as follows: the whole process presented in Figure 3 is performed on several values for the vocabulary size k. In case of planes combinations, the BoVFs obtained for every plane set are concatenated to form a final BoVF. The final BoVF representation dimension (i.e., the size of concatenated histograms coming from every plane set) was set between 60 and 900, in steps of 60.

For each vocabulary size, an extensive search for the SVM's penalty error C that would provide the higher recognition rate was done. A logarithmic scale between 10^{-10} and 10^{10} with 10 as a multiplicative step (logarithmic scale) was used for the C values. Every recognition rate for every k and C is measured in a 5-fold cross validation run. Once the best k and coarse C is found, an finer search of C values is performed, around the previous best one.

With the best k and best C at hand, ten new 5-fold cross validation runs are executed on the database, varying the

[1]A "frame set" is a set containing all the frames along a unique direction

Fig. 4: Snapshots for all actions in the Weizmann database [11]. From left to right, up to down: bending to the floor, jumping-jacking, jumping forward, jumping in place, running, walking laterally, jumping on one foot, walking, waving with one hand and with two hands.

random selection for the folds. The ten mean recognition rates found at these runs are averaged to compute the confidence intervals.

B. Results

Figure 5 shows the confidence intervals for the best recognition rates achieved with SURF points using different combinations of frames sets (at a 95% confidence level). The combinations are indicated in the y axis of this graph, while the recognition rates are indicated in the x axis.

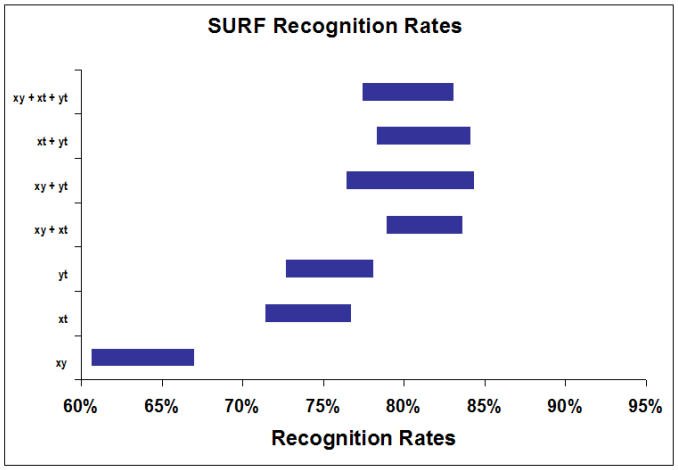


Fig. 5: Confidence intervals for the recognition rates obtained with SURF points gathered from different frames sets, at a confidence level of 95%.

From this graph, it is possible to see that just by using information from one of the spatio-temporal frames to build the video BoVF gives significant improvement on the recognition rate over the BoVF created from points detected on the original xy frames only ($\pm 11\%$ higher).

Also, the combination of the points coming from the xy (pure spatial) frames together with one of the spatio-temporal frames (xt OR yt) performs even better ($\pm 16\%$ more than the baseline).

Nevertheless, combining the points from all the frames together does not provide further improvement on the recognition rate, as it could be expected at a first sight. As can be seen from Figure 5, the recognition rates provided by the combinations $xy + xt$, $xy + yt$, $xt + yt$ and $xy + xt + yt$ have no statistically significant difference. This result indicates that while pure spatial and the spatio-temporal frames are complementary between them, spatio-temporal frames from different directions carry redundant information. Finally, the xy+yt combination provided the best results.

Figure 6 shows the results of the equivalent experiments with SIFT descriptors. As it can be noticed, these results are quite consistent with the SURF ones, including the fact that the recognition rates achieved by using points from xy and yt frames together are the best ones.

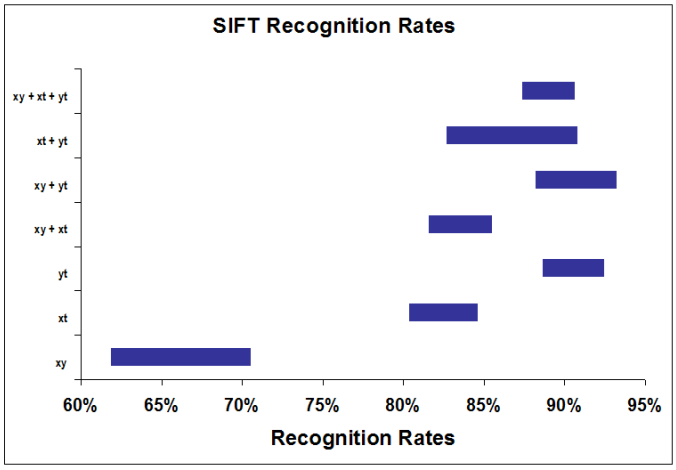


Fig. 6: Confidence intervals for the recognition rates obtained with SIFT points gathered from different frames sets, at a confidence level of 95%.

Since the Weizmann database is specifically focused on human actions, these results provide a strong indication that 2D interest points descriptors can indeed be used to capture dynamic information in videos, when applied to the spatio-temporal frames.

In Figure 7, the recognition rates for SURF and SIFT at several vocabulary sizes are presented. From this graph, it is possible to see that SIFT points consistently produce

Results on Weizmann DB	
Paper	**Rec. Rate**
[17]	72.8%
Ours (SURF)	$81 \pm 3\%$
[20]	82.6%
[19]	89.3%
[18]	90%
Ours (SIFT)	$91 \pm 3\%$

TABLE I: Comparing recognition rates of BoVF-based approaches applied to the Weizmann database. Some details of each comparing approach are provided in the text.

significantly higher recognition rates. This is probably due to the fact that SIFT selects more points, providing a denser sampling than SURF. This result contradicts the claims of SURF's authors since, at least in this scenario, SIFT performs better.

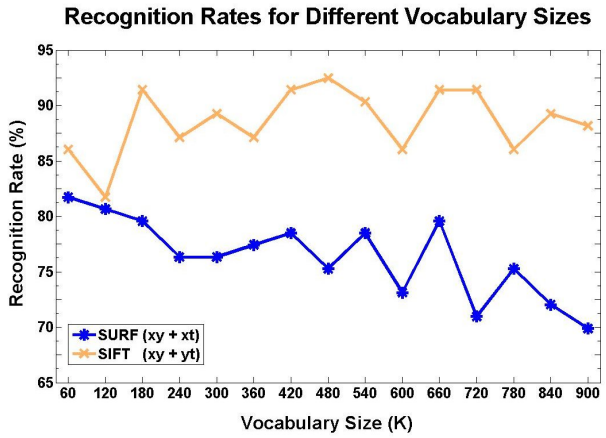


Fig. 7: Comparing results for SURF-based and SIFT-based approaches at various vocabulary sizes.

Finally, our best results are compared to other published results on the same database in Table I. Of course, this is quite a coarse comparison, because of the lack of a standard experimental protocol. Anyway, assuming confidence intervals similar to our ones, some discussion can be done on these numbers. Dollar's features (brightness gradients on space-time cuboids) are added with geometric information from a constellation model in [17], and a 3D extension for SIFT is proposed in [20]. These proposals are the more directly comparable to ours, since they deal with the temporal dimension with extensions of 2D descriptors, without further significative improvements to the basic BoVF. In those cases, our proposal produces better or equivalent recognition rates when SURF is applied, and much higher ones when SIFT is the choice for point selection and description.

The results achieved by [18] and [19] are considerably higher than our best SURF-based ones, but it is worth mentioning that in [18], Dollar's features are smoothed in varied scales, while we do not consider multi-scale features in our tests. Additionally, in [19], the model-free nature of a pure BoVF approach is lost, since in that work, Dollar's features are fused with features obtained from the actor's spatio-temporal volumes built from body silhouettes. By the other side, our plain BoVF implementation – with SIFT as the point selector/descriptor applied both to the traditional xy frames and to the yt spatio-temporal frames concatenated together – provides the highest average recognition rate.

V. Concluding Remarks

In this paper, we argue that 2D descriptors can be used to include dynamic information into a BoVF representation for videos, when applied to the spatio-temporal frames. To verify this assumption, a pure BoVF representation is applied to the recognition of human actions in videos from the Weizmann database, which has become a *de facto* standard for this task.

Both SURF and SIFT algorithms for interest point detection and description are compared, and our experimental results indicate that: a) 2D descriptors are indeed able to gather the dynamic information from videos, improving the recognition rates for concepts to which the dynamic aspect has a central role; b) contrary to the claims of SURF's authors, at least in this context, SIFT descriptors consistently present higher recognition rates; c) a plain BoVF representation, built on SIFT descriptors applied to collect information both from the traditional frames and the spatio-temporal frames provide state-of-the-art recognition rates for this database, even when compared to more complex approaches.

Future work includes the validation of these results on other actions databases and a deeper analysis on the issue of vocabulary formation for BoVFs representations, in search for a less empirical process for defining a visual vocabulary.

VI. Acknowledgments

The authors are thankful to CNPq, CAPES and FAPEMIG, Brazilian agencies, for the financial support to this work.

References

[1] Y.-G. Jiang, C.-W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *CIVR '07*, 2007, pp. 494–501.

[2] W. Hu, T. Tan, L. Wang, and S. Maybank, "A survey on visual surveillance of object motion and behaviors," *SMC-C*, vol. 34, no. 3, pp. 334–352, August 2004.

[3] K. S. Patwardhan and S. D. Roy, "Hand gesture modelling and recognition involving changing shapes and trajectories, using a predictive eigentracker," *Pattern Recogn. Lett.*, vol. 28, no. 3, pp. 329–334, 2007.

[4] G. Kosta and M. Benoit, "Group behavior recognition for gesture analysis," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 18, no. 2, pp. 211–222, Feb. 2008.

[5] A. Branzan Albu, T. Beugeling, N. Virji Babul, and C. Beach, "Analysis of irregularities in human actions with volumetric motion history images," in *Motion07*, 2007, pp. 16–16.

[6] J. K. Aggarwal and S. Park, "Human motion: Modeling and recognition of actions and interactions," in *3DPVT '04*. Washington, DC, USA: IEEE Computer Society, 2004, pp. 640–647.

[7] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *Comput. Vis. Image Underst.*, vol. 104, no. 2, pp. 90–126, 2006.

[8] R. Chellappa, A. K. Roy-Chowdhury, and S. K. Zhou, "Recognition of humans and their activities using video," *Synthesis Lectures on Image, Video, and Multimedia Processing*, vol. 1, no. 1, pp. 1–173, 2005.

[9] A. Mokhber, C. Achard, and M. Milgram, "Recognition of human behavior by space-time silhouette characterization," *Pattern Recogn. Lett.*, vol. 29, no. 1, pp. 81–89, 2008.

[10] R. A. Baeza-Yates and B. A. Ribeiro-Neto, *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.
[11] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *TPAMI*, vol. 29, no. 12, pp. 2247–2253, December 2007.
[12] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Speeded-up robust features (SURF)," *CVIU*, vol. 110, no. 3, pp. 346–359, 2008.
[13] D. Lowe, "Object recognition from local scale-invariant features," *ICCV '99*, vol. 2, pp. 1150–1157 vol.2, 1999.
[14] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *ICPR '04*, 2004, pp. III: 32–36.
[15] I. Laptev and T. Lindeberg, "Space-time interest points," in *ICCV '03*, 2003, pp. 432–439.
[16] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *ICCCN '05*. Washington, DC, USA: IEEE Computer Society, 2005, pp. 65–72.
[17] J. Niebles and F. Li, "A hierarchical model of shape and appearance for human action classification," in *CVPR '07*, 2007, pp. 1–8.
[18] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *IJCV*, vol. 79, no. 3, pp. 299–318, 2008.
[19] J. Liu, S. Ali, and M. Shah, "Recognizing human actions using multiple features," *CVPR '08*, pp. 1–8, June 2008.
[20] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *MULTIMEDIA '07*. New York, NY, USA: ACM, 2007, pp. 357–360.
[21] H. Ning, Y. Hu, and T. Huang, "Searching human behaviors using spatial-temporal words," in *Proceedings of the IEEE International Conference on Image Processing*, 2007, pp. 337–340.
[22] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," *CVPR '08*, pp. 1–8, June 2008.
[23] J. Liu and M. Shah, "Learning human actions via information maximization," in *CVPR '08*, June 2008.
[24] D. C. Lay, *Linear algebra and its applications*, 3rd ed. New York: Addison-Wesley, 2002.
[25] T. M. Mitchell, *Machine Learning*. New York: McGraw-Hill, 1997.
[26] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.