# Motion Silhouette-Based Real Time Action Recognition

Marlon F. de Alcântara, Thierry P. Moreira, and Helio Pedrini

Institute of Computing - University of Campinas
Campinas, SP, Brazil, 13083-852

**Abstract.** Most of the action recognition methods presented in the literature cannot be applied to real life situations. Some of them demand expensive feature extraction or classification processes, some require previous knowledge about starting and ending action times, others are just not scalable. In this paper, we present a real time action recognition method that uses information about the variation of the silhouette shape, which can be extracted and processed with little computational effort, and we apply a fast configuration of lightweight classifiers. The experiments are conducted on the Weizmann dataset and show that our method achieves the state-of-the-art accuracy in real time and can be scaled to work on different conditions and be applied several times simultaneously.

## 1 Introduction

The recent advances in technology have made computers faster, data storage cheaper and video capture more available. This provided extensive usage of applications of automatic human action recognition in video. They can be seen in surveillance systems, cell phones, cars and video games for various purposes. However, researchers face the efficiency-speed trade-off dilemma, seen in many computational problems, which hampers the implementation of real time solutions.

Over the last decade, numerous works have addressed video action recognition, aiming to achieve better classification rates. Eventually, the rates on some datasets have already reached around 100% – some examples are [5, 10, 13]. Hence, more recently, some works have consisted of making the techniques applicable to real life situations, even at the cost of reducing the classification rate. Some researches have addressed real time recognition [2, 8] and studied recognition of multiple actions simultaneously or in sequence [18]. The classification rate reduction in some recent works can be seen in Table 1.

Several methods presented in the literature, such as [10, 14], extract interest points from the video and describe them using solely appearance information. A Bag-of-Word approach is often used to unite all the local features, thus loosing their spatio-temporal distribution. These methods usually have a slow training phase and have limited application. Bregonzio et al. [1] developed a method in which the geometric information is preserved, obtaining better accuracy results,

J. Ruiz-Shulcloper and G. Sanniti di Baja (Eds.): CIARP 2013, Part II, LNCS 8259, pp. 471–478, 2013.

but still not solving these limitations. Another common approach is to use silhouettes; there are simple and fast ways to obtain them. One challenge of these approaches is to find a suitable form of representation; Yi and Krim [17] used an homotopy function to describe a space-time volume formed by silhouettes over time.

The work proposed by Guo et al. [6] have achieved an impressive success rate, however, it uses a dense set of feature vectors and covariance matrices. With some optimization, it can operate in real time, but is not scalable. Other methods, such as [15], work in real time, but have room for improvement in the correct classification rate and in the capability of recognizing multiple actions in space and time. Frequently, the reason why a method cannot be applied in real time is that the used features represent the entire action, therefore, the sequence must be acquired in order to call the classifier. Table 1 summarizes some related works, their accuracy and a short description of their techniques.

This paper proposes a lightweight action recognition method that is capable of identifying actions using only a small number of frames. The method is based on the shape variation of the motion silhouette, thus the features can be extracted on-the-fly and quickly be used to classify the action. For these characteristics, it can be readily applied to work with multiple simultaneous actions and actions in sequence.

**Table 1.** A summary of related works for the Weizmann dataset

| **Work** | **Weizmann rate (%)** | **Techniques** |
|---|---|---|
| Fathi and Mori (2008) [5] | 100 | Combination of low-level features with AdaBoost |
| Niebles, Wang and Fei-fei (2008) [10] | 90 | Bag-of-Words + pLSA |
| Sun, Bhen and Hauptmann (2009) [13] | 97.8 | Zernike moments + Bag-of-Words + SVM |
| Ta et al. (2010) [14] | 94.5 | Bag-of-Words + SVM |
| Hsieh, Huang and Tang (2011) [7] | 98.3 | Silhouette histogram in polar coordinates + Nearest Neighbor |
| Wang, Huang and Tan (2009) [15] | 93.3 | Optical Flow + AdaBoost |
| Bregonzio, Xiang and Gong (2012) [1] | 96.7 | Bag-of-Words + Clouds of Points + Multiple Kernel Learning |
| Junejo and Aghbari (2012) [8] | 88.6 | SAX + Nearest Neighbor |
| Zhang and Tao (2012) [18] | 93.9 | Slow Feature Analysis |
| Chaaroui and Climent-Pérez (2013) [2] | 90.3 | Silhouette clustering into key poses + Nearest Neighbor |
| Guo, Ishwar and Konrad (2013) [6] | 100 | Covariance matrix of spatio-temporal descriptors |

This paper is organized as follows. Section 2 defines the proposed methodology for this work. Section 3 presents and discusses some of the results obtained with the proposed method. Section 4 concludes the paper and includes some future work suggestions for improving the proposed method.

## 2 Methodology

The proposed method for identifying different actions is initialized with a video stream that contains an action, according to Figure 1. The first step is to acquire the motion silhouette by using the difference between consecutive frames; this step is fast to be applied and is responsible for the real time application.

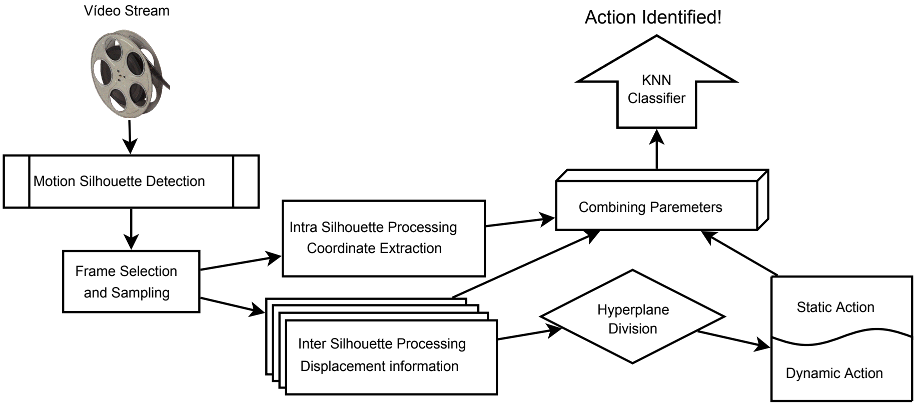
Vídeo Stream
Action Identified!
KNN Classifier
Motion Silhouette Detection
Combining Paremeters
Intra Silhouette Processing Coordinate Extraction
Frame Selection and Sampling
Hyperplane Division
Static Action
Dynamic Action
Inter Silhouette Processing Displacement information

**Fig. 1.** Diagram with the main stages of the proposed method

An action can be represented in the video stream with distinct number of frames. Some of these frames do not represent significant information to classify and can contain some outliers, due to the low robustness and fast speed method for acquiring the points. To overcome this weakness, the algorithm select frames to acquire the points with relevant information. Among the frames selected, a fixed number is sampled and will be used in the subsequent steps.

The extracted silhouettes are used in two distinct processes. The first one is the usage of a bounding box containing the entire silhouette. The bounding box contains some control points; the number of control points is parametrically defined and is equally divided into the four bounding box sides and equally spaced. The control points are used to choose the silhouette interest points; the selected points are those which the distance to each control point is the shortest; Figure 2 illustrates some control points and the selected points in a bounding box.

While the first step is applied to each silhouette separately, the second considers the silhouette point displacement in the same stream (Figure 3). The

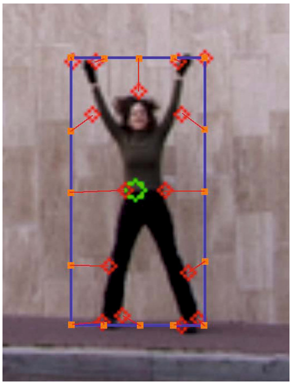

**Fig. 2.** Points of interest are chosen by their distance to the grid points

displacement is measured by using the Euclidean distance on the same points in different frames. Particularly, the displacement in the horizontal axis is used to differentiate static action from dynamic actions. While static actions start and end in the same place, dynamic actions start and end in different places. This information is used to create two hyperplanes, the first for static actions and the second for dynamic actions. Thus, any static action can be identified as a dynamic action and vice versa.

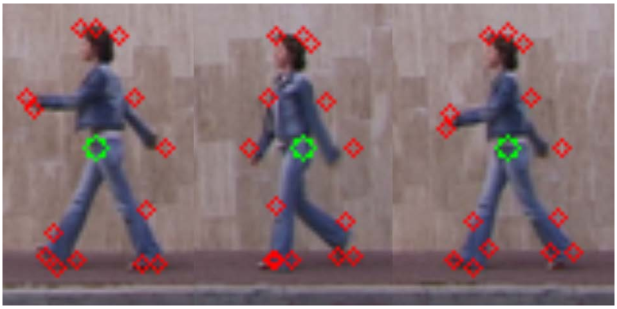

**Fig. 3.** Displacement of interest points in an action sequence

After the intra and inter silhouette process, the resulting parameters are combined into a unique descriptor and submitted to a classifier, which identifies the actions performed on the video stream. There are two viable classifier options: Support Vector Machine (SVM) [3] and $k$-Nearest Neighbors (KNN) [4].

SVM is originally a binary classifier. The training consists of finding a high-dimensional hyperplane that optimally separates the features of two classes. Multiclass classification is usually achieved by the use of several binary SVMs. Two common approaches are the *one-versus-all* and *one-versus-one*. In the first, each class is trained against all others together, and the classifier with the best output function gives the result. In the second, every class is trained against all others, one by one, and the result is given by a voting strategy.

KNN is a multiclass classifier. No training step is required, since the classification step uses the training vectors directly. It works by searching the space for $k$ nearest vectors from the testing instance. If $k$ is 1, it becomes a nearest neighbor classification.

The KNN classifier was chosen since it properly handles multiclasses and works very well in the coordinate system used in the proposed method, once the action classes tend to be organized into clusters. Also, uncorrelated classes usually do not weight in the classification, because only the surroundings of the test vector are considered.

## 3 Experimental Results

In this section, we evaluate our method on public datasets and present the results, as well as details of each method step, such as the tools and the parameters employed.

The experiments presented were conducted on the Weizmann human action dataset [16]. It consists of 10 classes of actions: *walking*, *bending*, *jumping jack*, *jumping*, *jumping in place*, *running*, *side walking*, *skipping*, *waving one hand* and *waving two hands*. Each action class is performed by 9 people, three of those classes have one person with two sequences recorded. Figure 4 shows some examples of actions from the dataset.

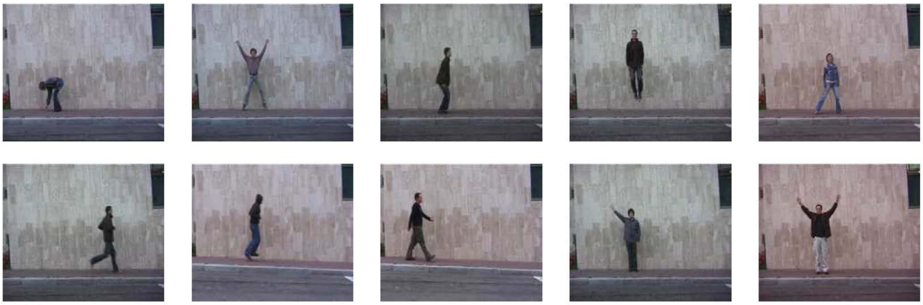

**Fig. 4.** Frames extracted from the Weizmann dataset [16]

To demonstrate the robustness of our method, we performed leave-one-out cross-reference tests. Table 2 shows the results obtained with the Weizmann dataset. It can be seen that 60% of the misclassification happens to the *skipping* class, as also reported by [2, 12], because it has a large intra class variation and is normally confused with *side walking* and *jumping*. The overall correct classification rate is 94.62%. This result shows that it is possible to achieve high rates of classification using simple descriptors, such as the proposed motion silhouette that allowed to perform real time data extraction and classification.

The processing time for a frame sequence is smaller than the video duration. The Weizmann dataset has 93 video sequences with an average duration of 2.45 seconds. The extraction of the features of each video took, in average, 0.135 seconds, and the average time to classify the videos is less than 1 millisecond. The ratio of the *average video duration* to the *average processing time* is *18.14*. The experiments were conducted in a 2.4GHz Intel i7 processor using no parallelism. It shows that our method works in real time with room for inserting improvements, without making it slow.

**Table 2.** Confusion matrix of the results for the Weizmann dataset

| | walk | bend | jack | jump | pjump | run | side | skip | wave1 | wave2 |
|---|---|---|---|---|---|---|---|---|---|---|
| walk | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| bend | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| jack | 0 | 0 | 0.77 | 0 | 0.22 | 0 | 0 | 0 | 0 | 0 |
| jump | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| pjump | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| run | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| side | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| skip | 0 | 0 | 0 | 0.2 | 0 | 0 | 0.1 | 0.7 | 0 | 0 |
| wave1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| wave2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

The classification results also show that not all video frames are necessary to identify the action correctly. After some tests using different number of frames, fifteen frames were used in the final algorithm. For tests using more than fifteen frames, no significant gain was observed.

Sixteen interest points were selected in the silhouettes (the number must be a power of two), resulting in a large number of final descriptor dimensions. The tests using more than sixteen points did not improve the identification process and the tests using less than sixteen showed a weak representation. To reduce the number of dimensions in the final descriptor, the PCA algorithm [11] was applied. Several dimensions were tested and the best value acquired was 30 to classify the video stream.

The KNN classifier allows a parameter to set a number of neighbors to be considered in the classification process. The best parameter observed in this case was one. It is because the used data set contains a short number of videos to train the classifier. In larger databases, this parameter could possibly be increased for a best classification.

Tests were also conducted on the KTH dataset [9], however, the method turned out to be ineffective on it since some videos have constant zooming and camera motion, which causes the detection of untrue displacement. This makes the displacement detection step that separates static from dynamic actions more difficult. The best recognition rate, reached by tunning the parameters, was 58.34%.

## 4 Conclusions

This paper introduced and discussed a new real time motion silhouette-based method for human action recognition. The most onerous part for any action recognition system is usually the descriptor extraction. In this work, we used a simple point selection that considers the relative point position for the control points fixed in a bounding box. This allows a fast silhouette representation and it is possible to recognize an action performed in a video stream correctly through a movement measure. The action sequence is described by the displacements of these points in time.

When using a silhouette-based algorithm, its robustness depends on the sampling adopted. In this step, the amount of points and which of them must be used are critical decisions for achieving high classification rates. To improve the results, our algorithm is capable of sampling a number of points based on the video resolution. For the Weizmann dataset, only sixteen control points were used, corresponding to sixteen interest point coordinates.

Unlike [1, 5, 6], the performance of our algorithm is more than necessary for real time requirements. Nevertheless, our classifier proved to be accurate and even better than other works of literature (Table 1) on the Weizmann dataset with an accuracy of 94.62%.

A motion-based algorithm is not indicated for videos containing camera motion, for instance the KTH dataset, since the method interprets a camera motion as movement, not segmenting correctly the action performed. A possible strategy is to use sophisticated tracking techniques such as [2], which, on the other hand, slow down the method performance.

The solution employed the proposed method is lightweight and easily scalable to work with multiple action instances on a single video sequence, because it is applied on each movement instance separately. Our approach achieves state-of-the-art 94.62% accuracy on the Weizmann dataset.

As future directions, we suggest to the proposed method for extracting and tracking silhouettes in the presence of more complex background, and also apply our method to different types of datasets, such as surveillance videos, videos with camera movements, or videos with other camera angles. It is possible that the descriptor developed in this work is suitable with other classifiers besides KNN.

**Acknowledgements.** The authors are grateful to FAPESP, CNPq and CAPES for the financial support.

## References

1. Bregonzio, M., Xiang, T., Gong, S.: Fusing Appearance and Distribution Information of Interest Points for Action Recognition. Pattern Recognition 45(3), 1220–1234 (2012)

2. Chaaraoui, A.A., Climent-Pérez, P., Flórez-Revuelta, F.: Silhouette-based Human Action Recognition using Sequences of Key Poses. Pattern Recognition Letters (2013)
3. Cortes, C., Vapnik, V.: Support-Vector Networks. Machine Learning 20(3), 273–297 (1995)
4. Cover, T., Hart, P.: Nearest Neighbor Pattern Classification. IEEE Transactions on Information Theory 13(1), 21–27 (1967)
5. Fathi, A., Mori, G.: Action Recognition by Learning Mid-Level Motion Features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
6. Guo, K., Ishwar, P., Konrad, J.: Action Recognition From Video Using Feature Covariance Matrices. IEEE Transactions on Image Processing 22(6), 2479–2494 (2013)
7. Hsieh, C., Huang, P., Tang, M.: The Recognition of Human Action Using Silhouette Histogram. In: Reynolds, M. (ed.) Proceedings of Australasian Computer Science Conference, vol. 113, pp. 11–16. ACS, Perth (2011)
8. Junejo, I.N., Aghbari, Z.A.: Using SAX Representation for Human Action Recognition. Journal of Visual Communication and Image Representation 23(6), 853–861 (2012)
9. KTH Royal Institute of Technology: KTH Action Dataset (2004), `http://www.nada.kth.se/cvap/actions/`
10. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. International Journal of Computer Vision 79(3), 299–318 (2008)
11. Pearson, K.: Principal Component Analysis. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 6(2), 559 (1901)
12. Saghafi, B., Rajan, D.: Human Action Recognition using Pose-based Discriminant Embedding. Image Communications 27(1), 96–111 (2012)
13. Sun, X., Chen, M., Hauptmann, A.: Action Recognition via Local Descriptors and Holistic Features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 58–65 (2009)
14. Ta, A.P., Wolf, C., Lavoue, G., Baskurt, A., Jolion, J.M.: Pairwise Features for Human Action Recognition. In: Proceedings of the 20th International Conference on Pattern Recognition, pp. 3224–3227. IEEE Computer Society, Washington, DC (2010)
15. Wang, S., Huang, K., Tan, T.: A Compact Optical Flowbased Motion Representation for Real-Time Action Recognition in Surveillance Scenes. In: Proceedings of the IEEE International Conference on Image Processing, Cairo, Egypt, pp. 1121–1124 (November 2009)
16. Weizmann Institute of Science: Weizmann Classification Database (2005), `http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html`
17. Yi, S., Krim, H.: Capturing Human Activity by a Curve. In: Proceedings of the IEEE International Conference on Image Processing, Cairo, Egypt, pp. 3561–3564 (November 2009)
18. Zhang, Z., Tao, D.: Slow feature analysis for human action recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 34(3), 436–450 (2012)