# Motion Primitives for Action Recognition

Fihl, Preben; Holte, Michael Boelstoft; Moeslund, Thomas B.

*Published in:*
GW 2007

*Publication date:*
2007

*Document Version*
Publisher's PDF, also known as Version of record

Link to publication from Aalborg University

*Citation for published version (APA):*
Fihl, P., Holte, M. B., & Moeslund, T. B. (2007). Motion Primitives for Action Recognition. In M. Sales Dias, & R. Jota (Eds.), GW 2007: The 7th International Workshop on Gesture in Human-Computer Interaction and Simulation (pp. 14-15). ADETTI/ISCTE.

# Motion Primitives for Action Recognition

P. Fihl, M.B. Holte and T.B. Moeslund

Laboratory of Computer Vision and Media Technology
Aalborg University, Denmark
Email: tbm@cvmt.dk

## 1 Introduction

Automatic recognition of human actions is a very active research area due to its numerous applications. A widely used approach is to do recognition directly on image data. These methods either represent an action by data from all frames constituting the action or by a number of smaller temporal sequences, e.g. atomic movements [1], dynamic instants [4], and key-frames [2]. Since image information can not allways be extracted reliably in every single frame the general idea is that approaches based on finding smaller units will be less sensitive compared to approaches based on an entire sequence of information.

In this paper we address action recognition using temporal instances (denoted primitives) that only represent a subset of the original sequence. We define primitives as instances with significant motion and an action is defined as a set of primitives. This approach allows for handling partly corrupted input sequences and does not require the lengths, the start point, nor the end point to be known, which is the case in many other systems. The focus of this work is five one-arm gestures representing the actions: *point right*, *raise the arm*, *move right*, *move left*, and *move closer*. The approach can with some modifications be generalized to body actions. Figure 1 illustrates the system.

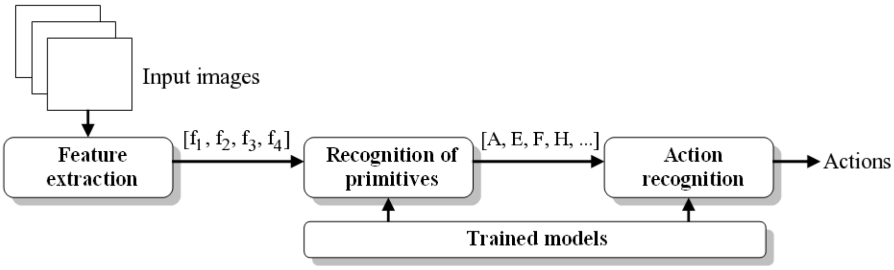

**Fig. 1.** System overview.

## 2 Action recognition system

Our primitives are based on motion and we extract this motion by generating double difference images. When doing arm gestures the double difference image will roughly speaking contain a ”motion-cloud”. Noise in the double difference image is reduced by a region growing approach in combination with a hysteresis threshold and the result is one connected motion-cloud. We model the motion-cloud compactly by an ellipse. The length and orientation of the axes of the ellipse are calculated from the Eigen-vectors and Eigen-values of the covariance matrix defined by the motion pixels. We use four scale and translation invariant features to represent this cloud. Each incoming frame is represented by the four extracted features and each feature vector is classified as a particular primitive or as noise based on a Mahalanobis classifier. This classification of a sequence can be viewed as a trajectory through the 4D feature space where the closest primitive (in terms of Mahalanobis distance) is found at each time-step. After processing a sequence the output will be a string with the same length as the input sequence and each letter of the string will represent a primitive (or noise).

During a training phase a string representation of each action to be recognized is learned. The task is now to compare each of the learned actions (strings) with the detected string. Since the learned strings and the detected strings (possibly including errors) will in general not have the same length, the standard pattern recognition methods will not suffice. We therefore apply the Edit Distance method [3], which can handle matching of strings of different lengths. The edit distance is a deterministic method but by changing the cost function of the method to represent the likelihoods of the primitives we make it a probabilistic method. Furthermore we normalize the edit distance to account for different lengths of the action-strings.

## 3 Results

The recognition rate was tested on a set of 550 action-sequences (11 persons doing 10 repetitions of each action). Two test scenarios were used. In the first scenario the start and stop times of the action were known. The sequences in the second scenario contained gesture-like motion both before and after the performance of the action so that the start and stop times of the action were unknown. Figure 2 shows the confusion matrices of the two tests. The overall recognition rates are $88.7\%$. and $85.5\%$, respectively.

| | 1. | 2. | 3. | 4. | 5. |
|---|---|---|---|---|---|
| 1. Point right | **100** | | | | |
| 2. Move left | **6.4** | **90.9** | | **2.7** | |
| 3. Move right | **5.5** | | **92.7** | **0.9** | **0.9** |
| 4. Move closer | | **2.7** | **1.8** | **70.9** | **23.6** |
| 5. Raise arm | | | | **10.9** | **89.1** |

(a) Known start and stop time.

| | 1. | 2. | 3. | 4. | 5. |
|---|---|---|---|---|---|
| 1. Point right | **99.1** | | **0.9** | | |
| 2. Move left | **9.1** | **90.0** | | **0.9** | |
| 3. Move right | **7.3** | | **90.0** | **2.7** | |
| 4. Move closer | **0.9** | **4.5** | **1.8** | **62.7** | **30.0** |
| 5. Raise arm | **1.8** | **1.8** | | **10.9** | **85.5** |

(b) Unknown start and stop time.

**Fig. 2.** The confusion matrices for the recognition rates (in percent) without added noise (a) and with added noise (b). Zero values have been left out to ease the overview of the confusion.

## References

1. L. Campbell and A. Bobick. Recognition of Human Body Motion Using Phase Space Constraints. In *International Conference on Computer Vision*, Cambridge, Massachusetts, 1995.
2. J. Gonzalez, J. Varona, F.X. Roca, and J.J. Villanueva. *aSpaces*: Action spaces for recognition and synthesis of human actions. In *AMDO*, pages 189–200, AMDO02, 2002.
3. V.I. Levenshtein. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Doklady Akademii Nauk SSSR*, 163(4):845–848, 1965.
4. C. Rao, A. Yilmaz, and M. Shah. View-Invariant Representation and Recognition of Actions. *Journal of Computer Vision*, 50(2):55 – 63, 2002.