Selecting Relevant Features for Human Motion Recognition

D. Gehrig and T. Schultz
Cognitive Systems Laboratory, Universität Karlsruhe (TH), Karlsruhe, Germany
{*dgehrig*|*tanja*}*@ira.uka.de*

Abstract

Recently, there is a growing interest in automatic recognition of human motion for applications, such as humanoid robots, human activity monitoring, and surveillance. In this paper we investigate motion recognition based on joint angle trajectories derived from marker-based video recordings. The goal of this paper is to improve the generalization and robustness of human motion recognition even if only limited amount of training data is available. We achieve this goal by significantly reducing the amount of input features. We leverage on recent studies in the area of neuroscience which indicate that human motions display only a few independent degrees of freedom (DOF). We examine which DOF are relevant for recognizing upper body human motions and to what extend the dimensionality of the feature vectors can be reduced in order to simplify the data acquisition and improve the robustness of the recognition process. Our final results indicate that careful selection of features proves to reduce the number of features by a factor of up to 3, while at the same time significantly improving the recognition performance.

1. Introduction

In the last years the interest in automatic recognition of human motion, e. g. in humanoid robot research has increased significantly. The overview article by Aggarwal and Park [2] describes the large variety of approaches, ranging from statistical modeling techniques such as Hidden Markov Models (HMMs) [8, 9] to biologically motivated recognizers [10]. Recently, research in neuroscience uncovers the process of motion generation in humans and animals, and indicates that human motions have only a few independent Degrees Of Freedom (DOF) and that synergies are used to handle the complexity of motions [5]. Park [6] applies these synergies in her approach for modeling human motion.

2. Human Motion Recognition

In this paper we study human motion recognition based on joint angle trajectories derived from marker-based video recordings. We describe how to select the most relevant features for upper body motion recognition. For feature selection we investigate various feature selection methods, and compare the resulting features based on recognition accuracy.

2.1 Scenario

We focus on human motions as they appear in a kitchen and food preparation scenario of CRC 588[1] [1], such as placing objects and pouring fluid into container and discriminate the following 10 motion sequences: rolling pastry (M1), pouring water (M2), planing apple (M3), grinding coffee (M4), sweeping (M5), grating apple (M6), stiring (M7), cutting cake (M8), cutting apple (M9), and pitching (M10). Each motion is described in terms of a sequence of motion units, such as fetching, maneuvering, and placing back an object. Many motions share the same motion units.

Human subjects were asked to perform these motions in a controlled setting (see Fig. 1). The objects were placed at fixed positions on the table. The subject stands at the table in a neutral position, i.e. both hands resting on the table. Starting from this neutral position the subject executes a predefined sequence of motion units, e. g. fetching an empty glas, fetching a bottle of water, pouring water into the glas, and putting both objects back. In between the motion units, the subject is going into neutral position.

2.2 Body Model

For reconstructing the motion data we use a human body model. We use a rigid upper body multibody

[1]This work has been supported by the Deutsche Forschungsgemeinschaft (DFG) within Collaborative Research Center 588 ”Humanoid Robots - Learning and Cooperating Multimodal Robots.

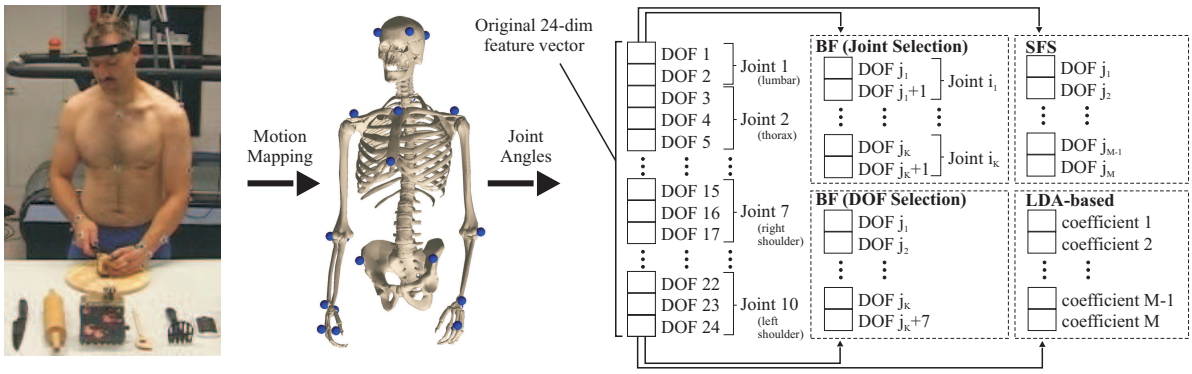

Figure 1. The data flow from data aquisition via preprocessing to feature selection.

model of the human skeleton with 24 DOF to approximate human motion. The joints are wrist (2 DOF), elbow (2 DOF) and shoulder (3 DOF) of both arms as well as joints at the lower (neck: 2 DOF) and upper (skull: 3 DOF) neck, and at the lower (lumbar: 2 DOF) and upper (thorax: 3 DOF) spine. The according DOF for wrist are for example wrist flexion and deviation. The joint angles of the 24 DOF are used as input features for our motion recognition system.

2.3 Motion Model

To recognize human motions, the feature vectors are fed into a 3-state left-to-right Hidden Markov Model which represents a motion unit. The three states describe the initial, middle, and final part of the motion unit. Each human motion is modeled as a concatenation of these motion units. In our experiments we discriminate 10 human motions as described above, consisting of 49 different motion units in total.

2.4 Data Acquisition

To capture the human motions, we attach reflecting markers to the subject's upper body, head and arms. Reflecting infra red light is simultaneously recorded with 10 Vicon cameras, which are arranged around the table. The Vicon system outputs 3-dimensional positions and labels of the markers. The resulting marker trajectories are used as input to an optimization-based motion mapping. This motion mapping determines the parameters of the kinematic model and calculates the related joint angle trajectories based on the kinematic body model by minimizing the distances between the marker positions in space and the body model. As a result, the motion mapping outputs per time step one feature vector consisting of the 24 joint angles of the kinematic body model.

2.5 Baseline Motion Recognition System

Our human motion recognition system features the one pass IBIS decoder [7], which is part of the Janus Recognition Toolkit (JRTk) [3]. For HMM motion unit model training we used about 600 recordings of human motions from a single subject. For model bootstrapping, we manually segmented about two thirds of these data into motion units. Each state of the motion unit left-to-right HMM has two equally likely transitions, one to the current state, and one to the next state. The emission probabilities are modeled by Gaussian mixtures, initialized by the K-Means algorithm based on the manually segmented data. The 24-dimensional feature vectors are normalized by subtracting the mean and normalizing the standard deviation to 1. HMM training was performed based on the standard forward-backward EM algorithm.

Two subsets of about 50 human motions each were held out as development (dev data) and test data, respectively. Recognition performance is reported throughout the paper in terms of motion unit recognition accuracy. Decoding was carried out as a time-synchronous beam search guided by a bigram model with a perplexity of 2.6 which describes the probabilities of two consecutive motion units. Large beams were applied to avoid pruning errors. This baseline system achieves 85.1 % recognition accuracy on the 49 motion units.

3. Experiments

The goal of our experiments is the selection of relevant features for recognizing human upper body movements in a kitchen environment. For the selection we implemented three methods, brute-force feature selection (BFS), sequential forward selection (SFS), and LDA-based feature selection (LDA). The three methods are compared using motion recognition accuracy.

3.1 Brute-Force Feature Selection (BFS)

We started the BFS method by investigating combinations of joints (see Fig. 1). When a joint was selected, each of its DOF was taken into consideration. Second, we refined our selection procedure by allowing combinations of DOF independent of the joints.

3.1.1 Selection by Joints

In this experiment, we trained a recognizer for each of the 1023 combinations possible with our 10 joints. This means we used feature vectors consisting of all DOF belonging to the selected joints.

The best recognition accuracy of 90.9 % was achieved when using right shoulder, right elbow, right wrist, and left elbow. The total amount of features was such reduced by a factor of roughly 3, outperforming the baseline system by 7 % (relative) on the dev data.

We examined the top 10 % of joint combinations. The two joints neck and skull seemed to be the least relevant for the recognition performance. Therefore, we discarded the worst joint (skull) in the following DOF experiments. Although the numbers indicated that the left shoulder might be less crucial as well, we did not remove this joint for generality reasons, as its relevance might depend on the handedness of the subject.

3.1.2 Selection by DOF

This experiment is based on the selection of individual DOF instead of the combination of DOF as predefined by the joints. Since this selection scheme allows for higher granularity, the results should provide better insights to which features are relevant for motion recognition performance. Since a motion can be performed either with the left or the right hand, we coupled the DOF of both arms such that the same DOF would be selected. For example, either the flexion for both, left and right arm, is included into the feature set, or both are discarded.

The large number of combinations is handled by separating the feature sets into two categories, a set of arm features and a set of spinal column and neck features. First, we examined all possible 127 combinations within the arm feature set while using all spine&neck features. Second, we examined the 127 combinations of arm features without using any spine&neck features. Then the same was done vice versa.

As can be seen in Table 1, the arm features outperform the spine&neck features. This is not too surprising given that the 10 human motions all focus on managing objects. However, it raises the question if additional spine&neck features on top of arm features could improve the recognition performance. We ranked the features according to frequency in the best 10 % of feature set combinations. Since the arm features are more important, we took the best five arm features and the best four spine&neck features and investigated all possible combinations. The best performance on the dev data was 92.5 % using thorax yaw, arm adduction, arm rotation, elbow flexion, and wrist flexion. To validate the result, we used the same DOF on the test data and got a recognition accuracy of 89.3 %, which is 5 % above the baseline system. The reduction to these nine features corresponds to a reduction factor of almost 3. As can be seen in Figure 3 the recognition accuracy has improved for almost all motion sequences. In total only 4 features are suffient to achieve the same accuracy as the baseline system (see Fig. 2).

selecting feature of	all features of	accuracy
spine&neck	arms	90.9 %
spine&neck	no arms	75.2 %
arms	spine&neck	88.2 %
arms	no spine&neck	90.4 %

Table 1. Recognition accuracys for the DOF selection experiment

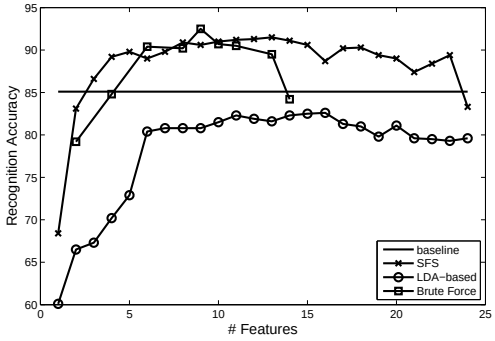

Figure 2. Performances over number of features for all selection methods

3.2 Sequential Forward Selection (SFS)

In this experiment we selected features based on sequential forward selection. Starting from a 1-dimensional feature vector (single DOF) we iteratively added one feature at a time in a greedy fashion until all 24 features were included. As criteria for adding a feature we used the recognition performance on the dev data. The test data was used for final evaluation. About 600 feature combinations were tested.

According to our results three features are already enough to outperform the baseline accuracy. The highest recognition accuracy was obtained with the follow-

ing 13 features: right shoulder (3 DOF) and supination (1 DOF), elbow flexion and wrist of both arms (6 DOF), thorax (2 DOF), and skull yaw (1 DOF). With these features we got a performance of 91.3% on the test data. As can be seen in Figure 3, the recognition results for most motion sequences are better than the baseline system.

3.3 LDA-based Feature Selection

We calculated a Linear Discriminant Analysis (LDA) [4] on training and dev data. The path alignments for the LDA were computed using a recognizer trained on the training data and initialized using the sequences of the training data, that have been manually segmented. We compared the LDA calculation based on two units, motion unit and HMM state and found that state-based LDA calculation achieves better results.

The best recognition accuracy has been achieved for 16 features of the LDA-transformed feature vector. Unfortunately the recognition accuracy of 81.8 % is below that of the baseline system. The LDA selection method improves over the baseline in only two human motion categories, ”apple planing” and ”apple grating”. Maybe these two highly confusable motions benefit from the discriminative nature of the LDA feature selection method.

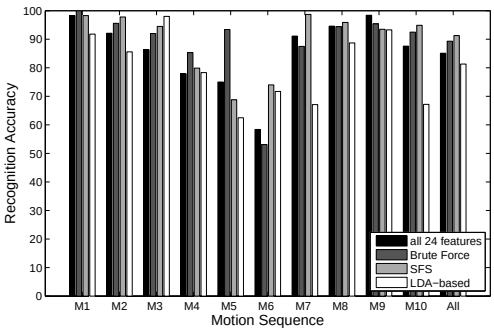

Figure 3. Breakdown of average performance for each motion using the optimized feature set per selection method

3.4 Comparison

Among the three feature selection methods, the BFS needs the most computation time and the LDA-based selection needs the least. While BFS relies on human expertise, SFS and LDA-based selection are solely data-based procedures. BFS and SFS both outperform the baseline system. While BFS is best on the dev data, SFS is best on the test data. Both selected almost the same features since all DOF that were used for the best BFS result are also part of the feature set for the best SFS.

4. Conclusion and Future Work

We investigated three methods for selecting relevant features for human motion recognition. BFS and SFS resulted in a significant increase of the recognition performance while reducing the number of DOF by a factor up to 3. In contrast, recognition performance of the LDA-based selection droped to about 82%, albeit the fastest method. The results indicate that joint angles of the arms, especially the dominant one, are more relevant than spine&neck in our scenario. The selections led to an increase in motion unit recognition rate from 85.1 % to over 90 %. Using a context free motion grammar we got a motion sequence recognition rate of 100 % (SFS).

For the future it might be interesting to investigate what distinguishes the relevant features from the ones that have been discarded in the experiments. The bad performance of some joint angles might be due to the joint itself and its importance for the motions. It might also be due to the feature extraction process, e. g. the optimization stability of the joint angle calculation for the particular joint.

References

[1] Collaborative research center 588 humanoid robots - learning and cooperating multimodal robots. http://www.sfb588.uni-karlsruhe.de/.

[2] J. Aggarwal and S. Park. Human motion: modeling and recognition of actions and interactions. *3DPVT 2004*, pages 640–647.

[3] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, and M. Westphal. The karlsruhe-verbmobil speech recognition engine. *ICASSP-97.*, 1:83–86 vol.1.

[4] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 2 edition, 1990.

[5] Y. Ivanenko, R. Poppele, and F. Lacquaniti. Five basic muscle activation pattern account for muscle activity during human locomotion, 2004.

[6] A.-N. Park, A. Mukovskiy, L. Omlor, and M. A. Giese. Self organized character animation based on learned synergies from foll-body motion capture data. In *Proceedings of the 2008 International Conference on Cognitive Systems*, pages 145–152, April 2008.

[7] H. Soltau, F. Metze, C. Füugen, and A. Waibel. A one-pass decoder based on polymorphic linguistic context assignment. *ASRU*, pages 214–217, 2001.

[8] T. Starner and A. Pentland. Real-time american sign language recognition from video using hidden markov models. In *ISCV ’95. Proceedings of the International Symposium on Computer Vision*, page 265, Washington, DC, USA, 1995. IEEE Computer Society.

[9] J. Yamato, J. Ohya, and K. Ishii. Recognizing human action in time-sequential images using hidden markov model. *CVPR*, pages 379–385, 1992.

[10] X. Yu and S. X. Yang. A study of motion recognition from video sequences. *Comput. Vis. Sci.*, 8(1):19–25, 2005.