9th IEEE-RAS International Conference on Humanoid Robots
December 7-10, 2009 Paris, France

HMM-based Human Motion Recognition with Optical Flow Data

Dirk Gehrig, Hildegard Kuehne, Annika Woerner and Tanja Schultz

Abstract— **Human motion recognition is traditionally approached by either recognizing basic motions from features derived from video input or by interpreting complex motions by applying a high-level hierarchy of motion primitives. The former method is usually limited to rather simple motions while the latter requires human expert knowledge to build up a suitable hierarchy. In this paper we propose a new approach that uses the strength of both methods while overcoming their respective limitations. Our approach is able to recognize the motion units within complex motion sequences. The recognition process applies Hidden Markov Models (HMM) based on features consisting of optical flow gradient histograms. For each primitive motion unit we train one HMM and then concatenate these primitive motion units to form complex motion sequences. Modeling sequences with HMMs allows for a very flexible combination of motion units into motion sequences. They can either be combined in a restrictive rule-based formulation using predefined grammars or be more flexibly combined using a statistical model of sequence probabilities. In this paper we are mainly interested in the comparison of the optical flow features with marker-based features, therefore we do not use a motion grammar.**

We apply our approach to 24 motion units forming five complex motion sequences as they appear in a real-world kitchen tasks. The results show that the proposed approach allows for a very fast low-level recognition of human motion units without the need for any complex reconstruction, post processing or pose estimation. Straight-forward characteristic flow fields in combination with HMM sequence modeling are sufficient to reliably recognize complex motions even with an unrestricted search. Our results show that this search already achieves 13.1 % recognition error rate. We compare HMM models based on the optical flow features to those derived from a marker-based system. Our recognition results indicate that optical flow features achieve a competitive performance.

I. INTRODUCTION

The recognition of human motion can be seen as one of the growing fields, perhaps even one of the key topics in human-robot interaction. Two different views on this problem have been established over the last few years. On the one hand the problem is considered from a computer vision point of view, dealing with the task of recognizing basic motions like pointing gestures or people walking or waving in front of a humanoid robot. As Moeslund [11] states the attention of these video-based approaches is usually limited to the recognition of simple motion units and avoids to go beyond this scope and to include any recognition of complex motion sequences.

On the other hand the recognition of human motion is seen as the problem of recognizing and interpreting complex motion sequences like pouring water into a bowl or setting a table. Here a lot of work has been done in the context of expressing complex motion sequences by a hierarchy of motion units and defining motion grammars for the recognition and generation of complex motion sequences. To provide a good basis the work in this area is mainly based on high-quality motion capture data which requires a complex and mostly static capturing setup.

The aim of this paper is to present a first step to overcome the gap between those two points of view. To do this, a video-based motion extraction method is combined with a higher level motion recognition system. The motion features are gained from optical flow information in the video sequence. To extract typical motion patterns, the optical flow directions are stored in histogramms, representing the amount of different motion directions in one frame.

These histograms are used as feature vectors which are fed into a 4-state left-to-right HMM representing a motion unit. A motion sequence is modeled as a concatenation of motion units. Modeling sequences with HMMs allows for a very flexible combination of motion units into motion sequences. They can either be combined in a restrictive rule-based formulation using predefined grammars or be more flexibly combined using a statistical model of sequence probabilities. The latter allows to recognize prior unseen sequences of motions.

As our focus is on housekeeping scenarios, five common kitchen tasks had been considered for recognition: pouring water into a bowl(1), grating an apple(2), stirring(3), cutting fruits(4) and mashing potatoes(5) as can be seen in Fig. 1. Each motion sequence consists of 6 - 9 motion units with 24 different motion units in total.

For the evaluation of the presented system, we captured each motion sequence simultaneously with a video camera of a humanoid robot head and with a marker-based motion capture system. So we were able to compare the performance of the video-based system with a marker-based motion recognition system.

The results show that the proposed approach allows for a very fast low-level recognition of human motion units without the need for any complex reconstruction, post processing or pose estimation. Instead simple, characteristic flow fields are enough to recognize a motion allowing also dealing with low-quality videos or situations, when the body is partly occluded. It is shown that it is possible to recognize an executed motion based only on the temporal distribution of motion gradients. This ability is mostly independent from

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

D. Gehrig, H. Kuehne, A. Woerner and T. Schultz are with the Institute for Anthropomatics, Universitaet Karlsruhe (TH), 76131 Karlsruhe, Germany {dgehrig|kuehne|woerner|tanja}@ira.uka.de

978-1-4244-4588-2/09/$25.00 ©2009 IEEE

any environmental appearances like e.g. temporal occlusions, but also from the visual representation of the object itself as e.g. its size, color or surface appearance.

II. RELATED WORK

As most methods dealing with the recognition of motion units and sequences are based on motion information to preserve temporal information it still shows diversity between methods based on video sequences and motion recognition based on HMMs or similar approaches as one can see looking at surveys dealing with motion recognition over the last few years [11] and [9]. As stated by [11] and [9] video based systems so far deal with the problems of initialization, model based tracking and pose estimation. But the information gained in this context is usually not used for higher level recognition processes.

An interesting approach in this context is the one of Lokman et al.[7]. Here 3D stereo coordinates of hand and head positions are tracked over time. The resulting motion information is used to train a Bayesian network. The concept of motion words is used to find a description related to the grammar of human speech and to improve recognition. Yet the dictionary is already more complex then the one of most recognition algorithms, the presented motions are still rather basic tasks like drinking, calling, shaking hands etc.

Motion recognition at unit based level based on video data is e. g. described by Efros et al. [1] and Danafar and Gheissari [5]. Efros et al. [1] use the optical flow field to estimate the action of players during a soccer match, a tennis match as well as ballet poses performed by male and female dancers. They track each person within a suitable window and try to match the extracted motion descriptors with preclassified motions using a k-nearest neighbor approach. Typical motions to recognize are e.g. running, walking as well as swinging in different directions for tennis and typical ballet moves like plie, releve etc. without any further motion recognition or analysis. Danafar and Gheissari [5] propose a motion recognition approach in the context of vision based surveillance using vertical and horizontal optical flow to recognize actions. They split the body into three parts, the head, the trunk and the legs calculating a vertical and horizontal optical flow histogram for each of these parts and evaluate the so extracted motion descriptors with an SVM. The evaluation is based on the KTH motion database containing videos of different motions like walking, running, jogging, hand clapping waving and boxing.

Another kind of input data is chosen by Bradski and Davis [3]. Here motion history gradients gained from silhouette images are used to recognize different full body movements like stretching, kneeling or walking. Other algorithms using motion information to recognize motions are proposed by Junejo et al. [8] and Rao et al. [4] evaluating motion properties like a self similarity matrix or motion trajectories of predefined body segments, e.g. the hand to recognize motion sequences like golf swings or opening a cabinet board.

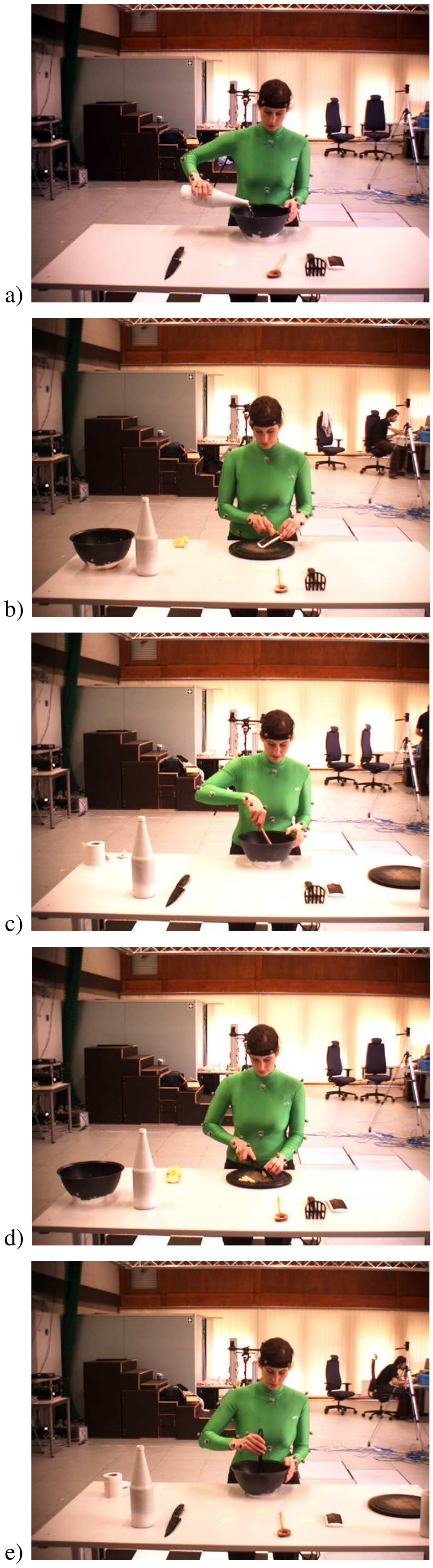

Fig. 1. Display of five complex human motion sequences in a kitchen scenario: pour water into a bowl (a), grate an apple (b), stir (c), cut fruits (d), and mash potatoes (e)

The recognition of more complex sequences is so far covered by Vogler and Metaxas using parallel HMMs to recognize American sign language [14] based on magnet tracking data, while Wilson and Bobick [15] propose parametric HMMs to recognize human gestures. Ryoo and Aggarwal [12] use HMMs for human motion recognition and combine it with recognized objects using DBNs in a hierarchical way.

III. SCENARIO

Our goal in this paper is to compare the motion recognition system without any markers on the observed person with a marker-based motion recognition system in a real world scenario. We use them for the recognition and concatenation of motion units into larger motion sequences. Our data were captured in a single session. A female test person performed each indicated task 20 times. The motion sequences were simultaneously recorded with a Vicon motion capture system and the camera system of a humanoid robot head. As video and marker data captures were taken at the same time, the markers are partly visible in the video images but are only used by the Vicon system.

Our scenario is part of the Collaborative Research Center 588 - Humanoid Robots. In this project the humanoid robot ARMAR has been developed. Its domain is the household of humans. Therefore our scenario takes place in a kitchen. The motion sequences usually comprise taking kitchen utensils from a table, working with them and putting them back to their places. The motion sequences had been manually segmented into motion units. These motion units had been defined beforehand using expert knowledge and are as follows:

Pouring water into a bowl: *Rest position - Take bowl - Take bottle - Pour - Put bottle back - Put bowl back - Rest position*

Grating an apple: *Rest position - Take grater - Take apple - Grate - Put apple back - Put grater back - Rest position*

Stirring: *Rest position - Take bowl - Take spoon - Stir - Put spoon back - Put bowl back - Rest position*

Cutting fruit: *Rest position - Take apple - Take knife - Grasp apple - Cut - Release apple - Put knife back - Put apple back - Rest position*

Mashing potatoes: *Rest position - Take bowl - Take masher - Mash - Put masher back - Put bowl back - Rest position*

If a cyclic motion unit is involved like stirring or grating, this motion unit is individually repeated 3-6 times per sequence. The test person was asked to do the motions as natural as possible.

IV. VIDEO BASED FEATURE ACQUISITION

For the image based representation of motion, histograms of optical flow directions weighted with their norm values are used. The videos of the motion sequences were captured with the left camera of a robot head, which was placed in front of the table. The robot cameras are dragonfly cameras with a resolution of 460 x 680 pixel and 30 fps. Every frame of the video sequence is represented by a histogram of its overall motion directions without any further local information.

The optical flow is computed using a pyramid version of the Lucas Kanade method, as described in [10], [2].

The weighted histogram for frame t is calculated from the optical flow OF of images (I_t, I_{t+1}).

$$OF(I_{t+\delta t}(x,y)) = I_t(x + u(\delta t), y + v(\delta t)) \tag{1}$$

The motion vector $(u(\delta t), v(\delta t))$ is used to calculated the resulting motion direction θ, indicated by an angle value from $[-\pi, \pi]$ and γ defining the motion intensity.

$$tan(\theta(u(\delta t), v(\delta t))) = \frac{v(\delta t)}{u(\delta t)} \tag{2}$$

$$\gamma(u(\delta t), v(\delta t)) = \sqrt{u(\delta t)^2 + v(\delta t)^2} \tag{3}$$

To simplify the notation the time index δt is presumed to be fix and omitted in the following. The elements for one bin of the histogram are calculated based on the motion angle θ. As the motion angle ranges from $[-\pi, \pi]$, the vector of elements for the k-th bin $h(k)$ of the histograms with n bins can be defined as:

$$\begin{aligned} h(k) = \{(u,v) | \theta(u,v) \geq \frac{(k2\pi)}{n} - \pi \cap \\ \theta(u,v) < \frac{((k+1)2\pi)}{n} - \pi\} \end{aligned} \tag{4}$$

The number of elements in $h(k)$ is indicated by $N(h(k))$ and the elements represent the coordinates (u,v) of the related optical flow vector. So the i-th element of $h(k)$ is defined as

$$h(k,i) = (u,v) \tag{5}$$

The k-th bin for the weighted histogram is calculated from the intensity of all elements in the vector as shown in

$$H(k) = \sum_{i=1}^{N(h(k))} \gamma(h(k,i)) \tag{6}$$

The histograms are sampled over time resulting in a multidimensional input vector for the HMMs. An example for the occurring optical flow vectors as well as for its weighted histogram can be seen in Fig. 2.

V. MARKER-BASED FEATURE ACQUISITION

Parallel to the video data acquisition the performed motions are recorded by a marker-based motion capture system (Vicon) in order to get a comparable baseline. For the marker based motion capture 10 Vicon cameras, which were arranged around the table, were used to capture the upper body motion of the test person with 100 fps. To capture the human motions 35 reflecting markers were attached to the subject's upper body, head and arms. The Vicon system outputs 3-dimensional positions and labels of the markers.

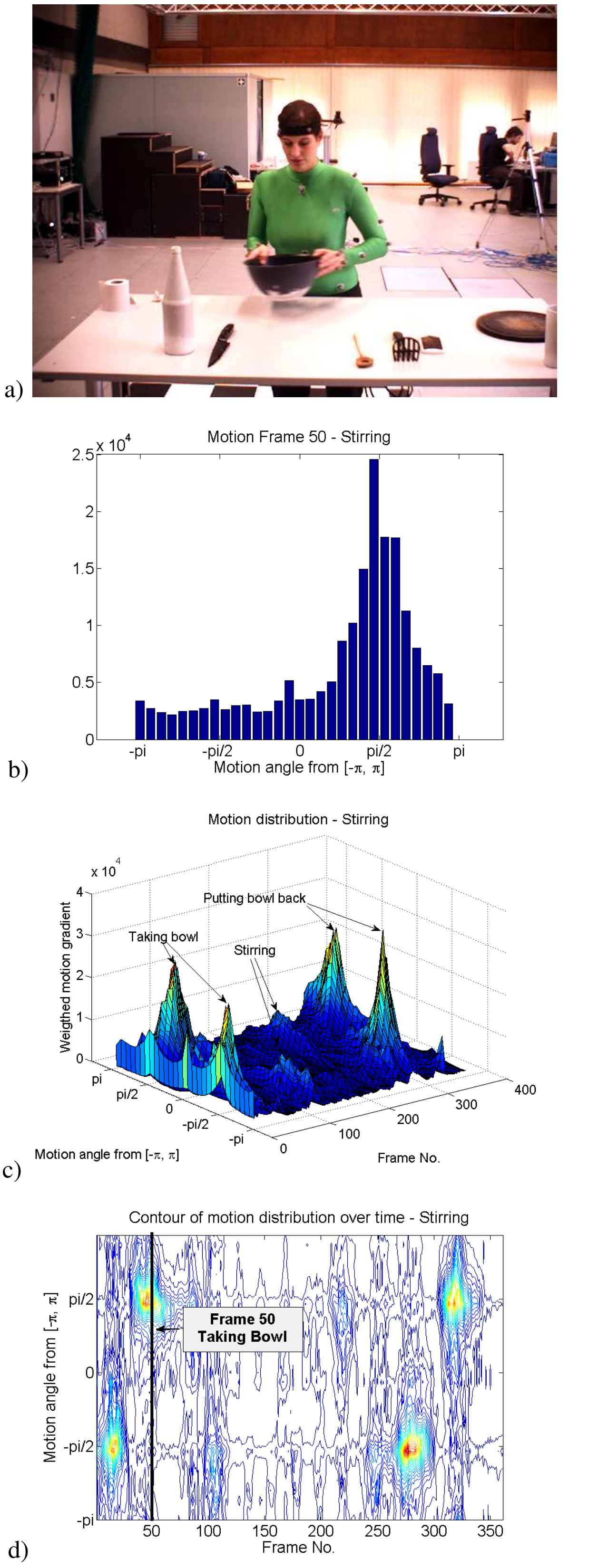

Fig. 2. Example for motion sequence *stirring*: (a) one frame of motion unit *taking bowl* as part of the motion sequence *stirring*, (b) motion gradient histogram for that particular frame, (c) distribution of optical flow gradients for the complete motion sequence, (d) motion distribution over time with two peaks indicating the beginning and the end of the motion sequence *stirring*.

The resulting marker trajectories are used as input to an optimization-based motion mapping. This motion mapping determines the parameters of a kinematic model and calculates the related joint angle trajectories based on a 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.

VI. EXPERIMENTAL RESULTS

A. *RECOGNITION SYSTEM*

Our human motion recognition system features the one pass IBIS decoder [13], which is part of the Janus Recognition Toolkit JRTk [6]. We used this toolkit to recognize human motions based on two types of input data: marker-based joint angles and video-based optical flow histograms. The following paragraphs describe the components of our system, i.e. the input features, the model topology, the model initialization, training, and optimization, as well as the decoding strategy. In order to ensure comparability, the marker-based recognition system differs from the video-based system only in the type of input feature, while all other system components are the same for both systems.

Feature vectors: The marker-based system uses a 24-dimensional feature vector as input, consisting of 24 joint angles from the upper body. The feature vectors of the video-based system contain the weighted optical flow histogram values for all motion angles. In our experiments we varied the number of dimensions between 15 and 45 to find optimal values. Results are given in the section below. The input feature vectors of both systems are normalized by mean subtraction and normalizing the standard deviation to 1. To compare the two data sets, both data were resampled to 20 fps.

Model units: Each of the 24 motion units is statistically modeled with a 4-state left-to-right Hidden Markov Model (HMM). In earlier experiments we found that 4 HMM-states work best in the described scenario. Each state of the left-to-right HMM has two equally likely transitions, one to the current state, and one to the next state. The emission probabilities of the HMM states are modeled by Gaussian mixtures. The number of Gaussians per mixture is optimized in cross-validation experiments described below. A motion sequence is then modeled as a sequential concatenation of these motion unit models. In total, we discriminate 5 types of human motion sequences as described above, consisting of the 24 different motion units.

Model initialization: To initialize the HMM models of the motion units, we manually segmented the marker-based data and the video-based data into the motion units. This was done separately for the two motion acquisition systems. The manually segmented data were then equally divided into four sections, and a Neural Gas algorithm was applied to initialize the corresponding HMM-state emission probabilities.

Model Training: For HMM motion unit model training and development we used 10-fold cross-validation on the 100 motion sequences. The training data of each fold contained

18 motion sequences of each of the 5 types of motion sequences. The remaining 2 motion sequences for each type were used for testing. For the experiments we varied the number of Gaussians between 1 and 64. For the video-based system we also varied the number of dimensions for the feature vectors. HMM training was performed featuring the Viterbi EM algorithm based on forced alignment on the unsegmented motion sequences.

Decoding: Decoding of the systems is carried out as a time-synchronous beam search. Large beams were applied to avoid pruning errors. We did not use a motion grammar since we are more interested in comparing the power of the two different types of features. Therefore, to guide the recognition process, we used a statistical zerogram model, which assigns the same prior probability (1/24) to each of the 24 motion units. Recognition performance is reported throughout the paper in terms of motion unit error rate as well as precision, recall and F-score, where F-score is defined as the harmonic mean of precision and recall:

$$F - score = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{7}$$

B. EVALUATION

First we investigated how many Gaussians per state give the best recognition results for a system with marker-based data. We used powers of 2 ranging from 1 to 64. Fig. 3 shows, that the best results are achieved, when using 16 Gaussians. The motion unit recognition error rate for this system is 17.6 %.

For the development of a system with the video-based data, we varied the number of Gaussians and the number of dimensions in the feature vectors. We again used powers of 2 ranging from 1 to 64 for the number of Gaussians. For the number of dimensions we tried 15, 30, 45. As can be seen in Fig. 3 we achieved a recognition error rate of 13.1 % when using 30 dimensional feature vectors and 16 Gaussians for each HMM state. Hence the video-based system outperforms the marker-based system. An overview of the two systems is given in Table I.

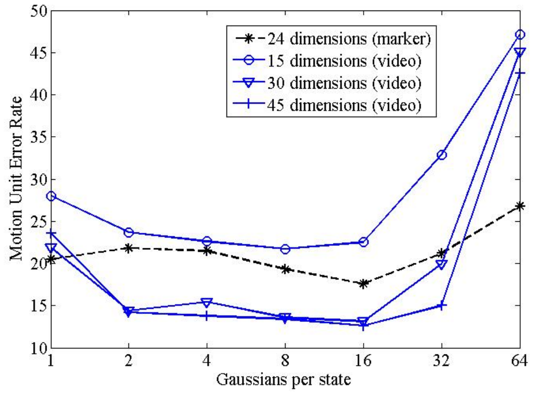

Fig. 3. Motion unit recognition error rates for marker-based and video-based systems

	Marker-based	Video-based
Feature vector dimension	24	30
Gaussians per state	16	16
States per motion unit	4	4
Motion unit error rate	17.6 %	13.1 %

TABLE I

COMPARISON OF BEST MARKER-BASED AND THE VIDEO-BASED SYSTEM

To get a better insight into the types of errors the two systems are producing, we also calculated precision, recall and F-score for each motion unit. Looking at the results in Fig. 4, we can see that cyclic motion units like *mash*, *grate* and *stir* have the worst F-score. Looking into the details it can be seen, that they all have a high precision and a low recall, which means that the majority of the existing cyclic motion units is not recognized. This is likely due to the following two reasons. Firstly, single cyclic motion units are usually the shortest motion units thus the HMM modeling might not be appropriate here. Secondly, due to the poor duration modeling, many deletion errors occur in cyclic motion units. These issues will be further investigated in future work.

When looking at the results in detail, we can also see, that all cyclic motions (*mash*, *grate*, *stir* and *cut*) have a worse recall with the marker-based data than with the video data. For the following motion units the F-score is worse with the marker-based system than with the video-based system: *rest position*, *take grater*, *put apple back* and *take apple*. The reasons for this differences have to be investigated in future works.

VII. CONCLUSIONS AND FUTURE WORKS

The presented approach allows a low-level recognition of human motion units and sequences without any complex reconstruction. It is shown that even similar motion units can be distinguished within motion sequences by using HMM-based motion unit recognition with optical flow motion gradient histograms. The recognition results are comparable to the error rate of marker-based motion recognition in our scenario.

Even if the systems perform reasonably well, the histogram based orientation of the current motion can result in problems that need to be solved in order to allow robust recognition. The training and testing of the presented approach is still limited to motions performed while facing the camera. To deal with the problem of people standing sideways to the camera, the test data as well as the training data for the HMMs will be extended to more than on frontal view comprising viewing angles of e.g. 45 and 90. Another way of handling this would be the usage of 3D motion information gained from stereo cameras, which we plan to investigate in the future.

A second restriction to overcome is that the person as well as the camera has to stay in one place to avoid ego motion to get reliable motion gradient histograms. Here a combination of the presented approach with person detection

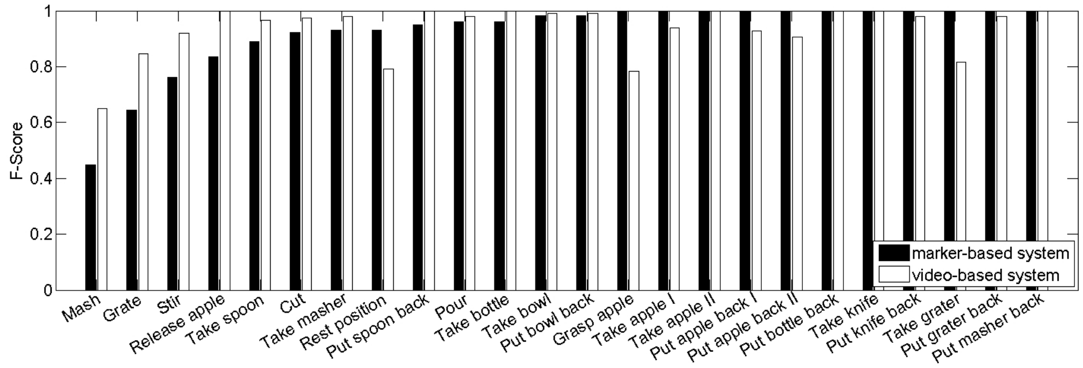

Fig. 4. Motion unit performance (F-Scores) for video-based system compared to marker-based system

and tracking algorithms could help to limit the calculation of optical flow gradient histograms to only those regions. This would not only result in a better performance, but it would also be possible to eliminate the persons global motion or camera motions, and so to recognize motions during a camera movement or when the person itself walks around. The focus on special regions could further allow to recognize the motion of more than one person. This would allow to recognize cooperative motion sequences between people like e.g. giving and taking kitchen utensils, one person stirring and one person pouring water in at the same time.

Further, more motion units and sequences have to be included testing the scalability of the presented approach as well as a more complex grammar allowing to model a larger and more flexible set of real world motions. The impact of sequence modeling on the performance has to be tested, i.e. manually created grammars and statistical models need to be explored. Besides, the investigation of person independent recognition is necessary.

VIII. ACKNOWLEDGMENTS

The authors would like to thank Andreas Fischer, Thorsten Stein and Hermann Schwameder from the institute for sports and sport science for the possibility to use the Vicon system and their assistance with the data acquisition. We would also like to thank our colleagues Martin Do and Pedram Azad for collecting the data with their robot head.

REFERENCES

[1] G. Mori A. A. Efros, A. C. Berg and J. Malik. Recognizing action at a distance. In *IEEE International Conference on Computer Vision*, pages 726–733, Nice, France, 2003.

[2] J. Y. Bouguet. Pyramidal implementation of the lucas kanade feature tracker: Description of the algorithm, 2002.

[3] G. R. Bradski and J. Davis. Motion segmentation and pose recognition with motion history gradients. In *Machine Vision and Applications*, pages 238–244, 2000.

[4] A. Yilmaz C. Rao and M. Shah. View-invariant representation and recognition of actions. *International Journal of Computer Vision*, 50(2):203–226, 2002.

[5] S. Danafar and N. Gheissari. Action recognition for surveillance applications using optic flow and svm. In *ACCV*, volume 2, pages 457–466, 2007.

[6] 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.

[7] Lokman J. Imai J. i. Kaneko M. Understanding human action in daily life scene based on action decomposition using dictionary terms and bayesian network. In *Second International Symposium on Universal Communication, 2008. ISUC '08.*, pages pp. 67 – 74, Dec. 2008.

[8] I. Laptev I. N. Junejo, E. Dexter and P. Pérez. Cross-view action recognition from temporal self-similarities. In *ECCV*, volume 2, pages 293–306, 2008.

[9] V. Krüger, D. Kragic, A. Ude, and C. Geib. The meaning of action: A review on action recognition and mapping. *Advanced Robotics*, 21(13):1473–1501, 2007.

[10] B. D. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision, 1981.

[11] Thomas B. Moeslund, Adrian Hilton, and Volker Krüger. A survey of advances in vision-based human motion capture and analysis. *Computer Vision and Image Understanding*, 104(2-3):90–126, 2006.

[12] M.S. Ryoo and J.K. Aggarwal. Hierarchical recognition of human activities interacting with objects. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, June 2007.

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

[14] Christian Vogler and Dimitris Metaxas. Parallel hidden markov models for american sign language recognition. In *In ICCV*, pages 116–122, 1999.

[15] Andrew D. Wilson and Aaron F. Bobick. Parametric hidden markov models for gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21:884–900, 1999.