GRAPHICAL MODELS FOR REAL-TIME CAPABLE GESTURE RECOGNITION

T. Rehrl, N. Theißing*, A. Bannat, J. Gast, D. Arsić, F. Wallhoff, G. Rigoll*

Human-Machine Communication
Department of Electrical Engineering and Information Technologies
Technische Universität München
Munich, Germany

C. Mayer, B. Radig*

Chair for Image Understanding and Knowledge-Based Systems
Computer Science Department
Technische Universität München
Munich, Germany

ABSTRACT

In everyday live head gestures such as head shaking or nodding and hand gestures like pointing gestures form important aspects of human-human interaction. Therefore, recent research considers integrating these intuitive communication cues into technical systems for improving and easing human-computer interaction. In this paper we present a vision-based system to recognize head gestures (nodding, shaking, neutral) and dynamic hand gestures (hand moving right/left/up/down, fist moving right/left) in real-time. The gestural input delivers a communication modality for a human-robot interaction scenario situated in an assistive household environment. The use of fast low-level image-feature extraction methods contributes to the real-time capability of the system and advanced classification approaches relying on Graphical Models provide high robustness. Graphical Models offer the possibility to group the input features in several sub-nodes resulting in a better classification than obtained via a traditional Hidden Markov Model classification. The applied grouping can regard interdependencies owing to, either physical constraints (like for the head gestures), or interrelations between shape and motion (like for the hand gestures).

Index Terms— gesture recognition, graphical models, real-time

1. INTRODUCTION

Human-machine interaction in general, as well as human-robot interaction in particular, constitute scenarios, which are currently of high interest in ongoing research efforts. Several research facilities have dedicated themselves towards the realization of a natural human-robot interaction. Due to the fact, that humans do not only rely on one interaction method, e.g. speech, human-robot interaction systems should be able to apprehend humans in a multi-modal way.

Owing to the fact that gestures are classical means of interaction in the humans' everyday life, many activities and efforts have been made by researchers to transfer these methods of communication into present technical systems. Head gestures are pleasant means to show agreement or disagreement [1]. Nonetheless, it is also imaginable to use head gestures for controlling operations, like in [2] for document browsing.

Hand gestures are very widespread in human-human communications, besides they deliver a large variety of expression possibilities ranging form pointing gestures over gesticulation towards language-like gestures up to sign language. A classical paper utilizing hand gesture recognition [3] resembles partially our approach, however, we have extended the Hidden Markov Model [4] based approach by relying on a more general classification method – directed Graphical Models. Our area of application is situated in a typical living environment and thus the detection of the human hand is more difficult and intricate. Robust hand detection in cluttered surroundings is also a problem for the approach of [5].

In general, more and more research has been conducted to cover the demands and requirements of the emerging research field of ambient assistive living, where objective is to assist elderly and disabled people. In [6] a remote control is introduced that utilizes ten predefined hand gestural commands to control a device selected via a pointing gesture. In [7] a system is presented that is capable to recognize hand gestures for human-robot interaction context.

The rest of this paper is organized as follows: In Section 2, the real-time capable framework for the gesture interaction analysis is presented. Section 3 considers the input features for the head and hand gesture Graphical Model-based classification process, which is delineated in Section 4. In Section 5 the obtained results with the presented approaches are regarded. The paper closes with a summary and an outlook over the next planned working steps.

2. FRAMEWORK

In the subsequent lines, we will have a closer look on the underlying framework of the dialog system facilitating the real-time capable head and hand gesture recognition. As it can been seen in [8], the Real-Time Data Base (RTDB) is capable of dealing with large amount of data input streams, having diverse features (i.e. data rate, packet losses, etc.). It is a shared memory implementation including a data buffer recording varying data for a certain time period and making these data available for different modules. Besides of a low computational processing overhead of the RTDB, different modules can process the same data originating from one input source without any blocking effects.

In general there are two concepts: The so-called "RTDB-Writer-Module" is necessary for making different data available in the RTDB memory. The counterpart the so-called "RTDB-Reader-Module" constitutes the necessary connection process handling, that a software-module can retrieve data from the RTDB memory.

For the processing of the gestures one RTDB-Writer is sufficient delivering the raw image data obtained from a webcam. This input stream is used by two different RTDB-Readers: head gesture feature extraction and hand gesture feature extraction. The obtained results from both feature extraction processes are stored in the RTDB again, where the recognition processes (Graphical Model-based or Hidden Markov Model-based) can classify the head or hand gestures. From the current frame f_t, a certain history length T (i.e. $T = 10$) of feature results $(f_{t-T}, \cdots, f_t)$ is handed over to the classification processes.

*These authors contributed equally to the work presented in this paper.

This work has partially been supported by the DFG excellence initiative research cluster *Cognition for Technical Systems – CoTeSys, www.cotesys.org*.

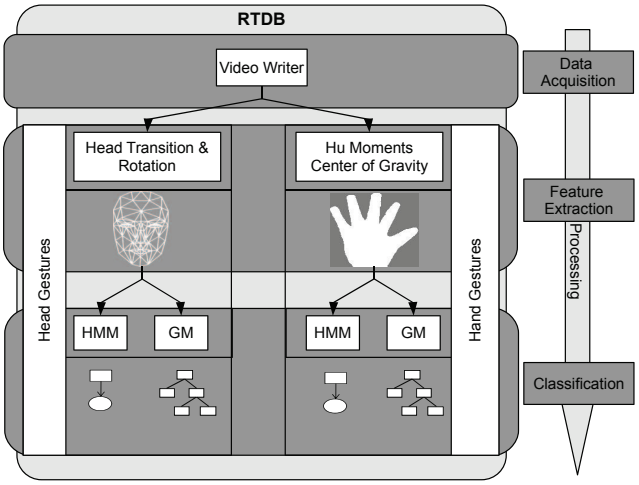

Fig. 1. Processing framework with the RTDB as a backbone.

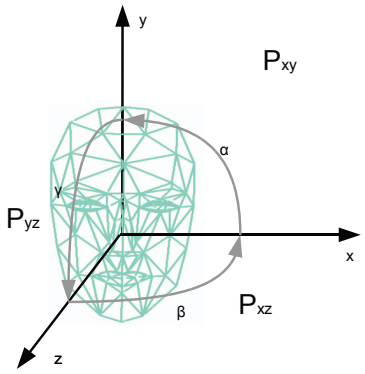

Fig. 2. The face model presented in the used coordinate system.

3. FEATURES

Several features are thinkable as the input information for the classification processes for the dynamic head and hand gestures. For ensuring a real-time capable behavior of the gesture recognition framework, we rely on features, which can be extracted from the image data fast to meet the real-time requirements. Different feature sets are used for the classification of the head and hand gestures. First the features for the head gestures will be considered, afterwards, the features for the hand gestures will be introduced.

3.1. Head Gestures

As a starting point for the localization of human heads in the video image obtained from the RTDB memory, we apply an implementation of the well-known Viola Jones approach for detecting human faces [9]. If more than one human is verified within the face detection process, we apply the following assumption: The service robot interacts with the human having the largest face in the current image.

Applying a sophisticated three dimensional model fitting strategy [10] and the Candide-III face model [11], we obtain an abstraction of the human face (see Figure 2), which yields to a characterizing parameter set. From this set we extract a parameter vector Θ_t consisting of the face transition (px_t, py_t, pz_t) and the face rotation $(\alpha_t, \beta_t, \gamma_t)$, yielding $\Theta_t = (px_t, py_t, pz_t, \alpha_t, \beta_t, \gamma_t)^T$.

For the classification only the temporal changes $\Delta\Theta_t$ are considered, given by:

$$\Delta\Theta_t = \Theta_t - \Theta_{t-1} \quad \forall t \subseteq \{1, ..., n\}, \quad \Theta_{t=0} = \mathbf{0} \qquad (1)$$

The obtained feature vector is submitted to the RTDB, where the classification process can readily access these vectors.

3.2. Hand Gestures

Similar to the approach mentioned in [12], we adapt the gained skin color model with regard to the obtained face image. This step improves the detection of the human hands and reduces the failure rate. In a defined region with regard to the determined head the hand gestures can occur. In this region the following rule is applied: All possible face candidates are rejected, and only the largest skin color blob is used for the hand contour image $Im(x, y, t)$. From the obtained hand contour image $Im(x, y, t)$ for time instance t, the center of gravity $\Lambda_t = (cx_t, cy_t)^T$ is determined by utilizing the image moments given by:

$$\mu_{pq} = \sum_x \sum_y (x - \bar{x})^p (y - \bar{y})^q Im(x, y, t).$$

Thus, the center of gravity is given by:

$$cx_t = \frac{\mu_{10}}{\mu_{00}} \qquad cy_t = \frac{\mu_{01}}{\mu_{00}}.$$

3.2.1. *Position related features*

The center of gravity constitutes the foundation for the position-related features. For the classification only the temporal changes $\Delta\Lambda_t$ are considered, given by:

$$\Delta\Lambda_t = \Lambda_t - \Lambda_{t-1} \quad \forall t \subseteq \{1, ..., n\}, \quad \Lambda_{t=0} = \mathbf{0} \qquad (2)$$

3.2.2. *Shape related features*

For the feature extraction process considering the shape of the human hand, we rely on features that are invariant to scale, translation, and rotation as well. Therefore, the Hu moments [13] are chosen from a set of possible candidates, since the following reason: The computation of the Hu moments can be performed very quickly, and thus, they are advantageous for real-time constraints. Thus the vector $\xi_t = (hu1_t, hu2_t, hu3_t, hu4_t, hu5_t, hu6_t, hu7_t)^T$ comprises the shape related features.

Both, the position related feature vector Λ_t and the shape related vector ξ_t are comprised in a new vector ζ_t, which is handed over into the RTDB, where classification process can easily access this nine features comprising vector ζ_t.

4. GRAPHICAL MODELS

Graphical Models (GMs) [14] are applied in many areas of research, since they provide descriptive and illustrative way to depict problems regarding control theory, computer science, pattern recognition, etc. In general, the GMs combine probability theory and

graph theory portraying the interdependences between different random variables (RVs). In this paper we consider only directed GMs, also known as Bayesian Networks (BNs). When BNs model time series data, they are called Dynamic Bayesian Networks (DBNs), which we apply to set up models for the gesture recognition. Hidden Markov Models (HMMs) are a sub-class of DBNs, where observations are dependent on an unobservable variable, referred as *hidden state*. For the realization of our GMs for classifying head and hand gestures, we used the Graphical Model Toolkit [15], whereas for the realization of the HMMs the Hidden Markov Toolkit [16] was used. In the following two sections the GMs constituting the gesture classification method will be presented. First, the GM for the head gesture classification process is considered, afterwards, the GM for hand gesture classification process is delineated.

4.1. GM for Head Gestures

The GM, shown in Figure 3, is composed of four discrete nodes, indicated via an unshaded rectangular form. The node labeled gesture represents the three different classes (nodding, shaking, and neutral). The remaining three discrete sub-nodes represent motion in two planes. The input for the three discrete nodes (State $P_{xy}P_{xz}$, State $P_{xy}P_{yz}$ State $P_{xz}P_{xy}$) are given by the temporal changes of the head transition ($\Delta px_t, \Delta py_t, \Delta pz_t$) and temporal changes of the head rotation ($\Delta\alpha_t, \Delta\beta_t, \Delta\gamma_t$). The input for a discrete sub-node is composed of two observations (shaded circles), one temporal change of a head transition and one temporal change of a head rotation. The observations were grouped, like they are combined via a gesture, e.g., for head shaking, the rotation angle of the head changes primarily in the β component and the position is altered in the px direction.

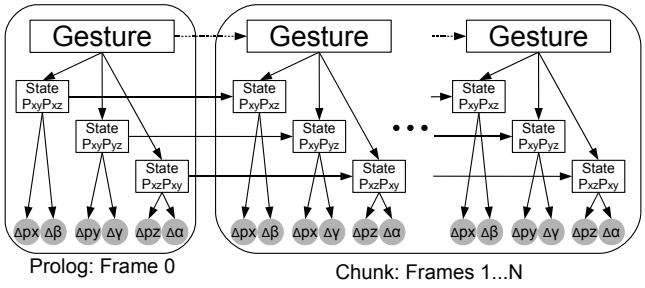

Fig. 3. Graphical Model for the head gesture classification.

4.2. GM for Hand Gestures

The GM, presented in Figure 4, is applied for the hand gesture recognition. Similar to the GM for the head gestures, the node labeled gestures represents the six different gesture classes: *fist to right*, *fist to left*, *hand to right*, *hand to left*, *hand up*, and *hand down*. One important characteristic of this GM is the different amount of nodes in the prolog and chunk. The prolog is composed of two subnodes: $StateMotion$, and $StateShape$. The input for the $StateShape$-Node are the seven hu moments ($hu_1, hu_2, hu_3, hu_4, hu_5, hu_6, hu_7$), which are used to determine the shape of the gesture (fist, hand). Due to the fact, that the shape does not change during the motion, it is sufficient and advantageous to determine the shape only in the first frame. The motion in the remaining frames $(1, \cdots, N)$ introduces noise, which can decline the recognition of the shape. The $StateMotion$- node uses the temporal changes of the center of gravity ($\Delta cx_t, \Delta cy_t$) for all frames $(1, \cdots, N)$ to determine the motion: move to right, move to left, move up, and move down.

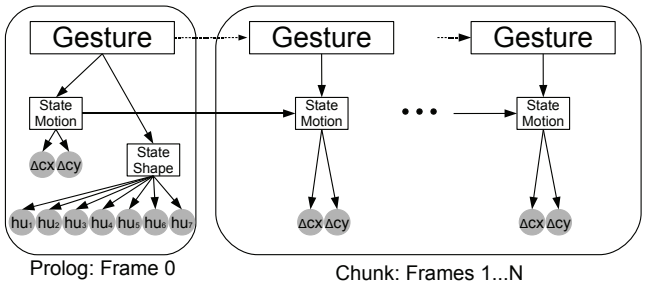

Fig. 4. Graphical Model for the hand gesture classification.

5. EXPERIMENTS AND EVALUATION

The trained GMs for head and hand gestures were evaluated with a five-fold cross validation. Therefore, the recorded data – 20 samples per class for each head/hand gesture – was split in five non-overlapping parts. Four parts were taken for training and the remaining fifth part was taken for testing. The process was iterated five times and the average of the resulting accuracy values was inspected. In addition, we compared the presented GM-based approach with a classical left-right HMM searching significant differences in the two recognition systems (only the best HMM-based results as well as the best GM-based results are presented in the following). First, the results regarding the head gestures will be investigated, afterwards the obtained results from the hand gestures.

For the significance test [17], we applied the following formulas to the determine a test statistic Z using the two recognition rates R_1 for GM-based classification system M_1 and R_2 for the HMM-based classification system M_2, respectively. To obtain a result for the significance test, a random variable $X(t)$ is introduced having the following properties:

$$X(t) = \begin{cases} 1 & M(1) \text{ correct and } M(2) \text{ wrong} \\ 0 & M(1) \text{ and } M(2) \text{ both correct or wrong} \\ -1 & M(1) \text{ wrong and } M(2) \text{ correct} \end{cases}$$

The test statistic Z applies the mean μ_x and the variance σ_x^2 of the random variable $X(t)$ given by

$$Z = \frac{\mu_x}{\sqrt{\frac{\sigma_x^2}{|T|}}}, \tag{3}$$

where $|T|$ is the number of samples: $|T| = 60$ for the head gestures and $|T| = 120$ for hand gestures, respectively.

The results for the head gestures can be seen in Table 1. For the GM-based system M_1^{head} a recognition rate $R_1^{head} = 90.00\%$ is obtained, whereas the 8-state HMM-based recognition system M_2^{head} has a recognition rate $R_2^{head} = 81.67\%$. The difference in the recognition rate has a significance given by the significane level of $Z, p(Z) = 97.4\%$.

The results for the hand gestures can be seen in Table 2. For the GM-based system M_1^{hand} a recognition rate $R_1^{hand} = 99.17\%$ is obtained, whereas the 8-state HMM-based recognition system M_2^{hand} has a recognition rate $R_2^{hand} = 94.17\%$. The difference in the recognition rate has a high significance given by the significane level of $Z, p(Z) = 99.4\%$.

Classification result using GM-based approach			
Classified	Sequence Label		
As	Shaking	Neutral	Nodding
Shaking	100%	0%	0%
Neutral	0%	95%	5%
Nodding	10%	15%	75%

Classification result using HMM-based approach			
Classified	Sequence Label		
As	Shaking	Neutral	Nodding
Shaking	100%	0%	0%
Neutral	0%	90%	10%
Nodding	10%	35%	55%

Table 1. This table presents recognition rates of our GM-based and HMM-based approaches for the head gestures. The results are obtained from a five-fold cross validation.

Classification result using GM-based approach						
	Sequence Label					
Classified as	Fist right	Fist left	Hand right	Hand left	Hand up	Hand down
Fist right	100%	0%	0%	0%	0%	0%
Fist left	0%	100%	0%	0%	0%	0%
Hand right	0%	0%	100%	0%	0%	0%
Hand left	0%	0%	0%	95%	0%	5%
Hand up	0%	0%	0%	0%	100%	0%
Hand down	0%	0%	0%	0%	0%	100%

Classification result using HMM-based approach						
	Sequence Label					
Classified as	Fist right	Fist left	Hand right	Hand left	Hand up	Hand down
Fist right	100%	0%	0%	0%	0%	0%
Fist left	0%	90%	0%	10%	0%	0%
Hand right	10%	0%	90%	0%	0%	0%
Hand left	0%	5%	0%	90%	0%	5%
Hand up	0%	0%	0%	0%	100%	0%
Hand down	0%	5%	0%	0%	0%	95%

Table 2. This table presents recognition rates of our GM-based and HMM-based approaches for the hand gestures. The results are obtained from a five-fold cross validation.

6. CONCLUSION AND FUTURE WORK

We introduced a real-time capable framework for the recognition of multiple gesture. GMs are applied to classify head gestures as well as hand gestures. The system has not reached its ultimate state, nonetheless, the GM-based approaches proved their capabilities by significantly outperforming the classical right-left HMMs and still preserving time constraints. Subsequent steps will tackle the robustness of extracted features by integrating additional depth information from time-of-flight cameras to improve hand segmentation and gesture classification as well. Furthermore, the gesture vocabulary will be extended. Finally, a fusion of results originating from different classification methods is imaginable.

7. REFERENCES

[1] C. Morimoto, Y. Yacoob, and L. Davis, "Recognition of head gestures using hidden markov models," in *In Proceeding of ICPR*, 1996, pp. 461–465.

[2] L.-P. Morency and T. Darrell, "Head gesture recognition in intelligent interfaces: the role of context in improving recognition," in *Proceedings of the 11th international conference on Intelligent user interfaces*, 2006, pp. 32–38.

[3] T. Starner, J. Weaver, and A. Pentland, "Real-time american sign language recognition from video using hidden markov models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 1371–1375, 1998.

[4] L. R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," in *Proceedings of the IEEE 77*, 1989.

[5] M. Kölsch and M. Turk, "Fast 2d hand tracking with flocks of features and multi-cue integration," in *In IEEE Workshop on Real-Time Vision for Human-Computer Interaction (at CVPR)*, 2004, p. 158.

[6] J. Do, J. Jung, S.H. Jung, H. Jang, and Z. Bien, "Advanced soft remote control system using hand gesture," in *5th Mexican International Conference on Artificial Intelligence, Apizaco, Mexico*, November 2006, pp. 215–220.

[7] Md. Hasanuzzaman, T. Zhang, V. Ampornaramveth, H. Gotoda, Y. Shirai, and H. Ueno, "Adaptive visual gesture recognition for human-robot interaction using a knowledge-based software platform," *Robot. Auton. Syst.*, vol. 55, no. 8, pp. 643–657, 2007.

[8] M. Goebl and G. Färber, "A real-time-capable hard- and software architecture for joint image and knowledge processing in cognitive automobiles," *Intelligent Vehicles Symposium*, pp. 737 – 740, June 2007.

[9] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Conference on Computer Vision and Pattern Recognition*, 2001, pp. 511–518.

[10] M. Wimmer, C. Mayer, F. Stulp, and B. Radig, "Estimating natural activity by fitting 3D models via learned objective functions," in *Workshop on Vision, Modeling, and Visualization (VMV)*, Saarbrücken, Germany, November 2007, vol. 1, pp. 233–241.

[11] J. Ahlberg, "Candide-3 – an updated parameterized face," Tech. Rep. LiTH-ISY-R-2326, Linköping University, Sweden, 2001.

[12] K. Nickel and R. Stiefelhagen, "Visual recognition of pointing gestures for human-robot interaction," 2007, pp. 1875–1884, Elsevier.

[13] M.K. Hu, "Visual pattern recognition by moment invariants," *IRE Transaction on Information Theory*, pp. 179–187, 1963.

[14] M. I. Jordan, Ed., *Learning in graphical models*, MIT Press, Cambridge, MA, USA, 1999.

[15] J. Bilmes, "GMTK: The graphical models toolkit," 2002.

[16] J. Odell, D. Ollason, P. Woodland, S. Young, and J. Jansen, *The HTK Book for HTK V2.0*, Cambridge University Press, Cambridge, UK, 1995.

[17] S. Günter, "Vergleich von Erkennungsmethoden," Tech. Rep. IAM-04-001, Institut für Informatik und angewandte Mathematik – Universität Bern, 2004.