Did I Get It Right: Head Gestures Analysis for Human-Machine Interactions

Jürgen Gast[1], Alexander Bannat[1], Tobias Rehrl[1], Gerhard Rigoll[1], Frank Wallhoff[1], Christoph Mayer[2], and Bernd Radig[2]

[1] Department of Electrical Engineering and Information Technology, Institute for Human-Machine Communication
[2] Computer Science Department, Chair for Image Understanding and Knowledge-Based Systems
Technische Universität München
80290 Munich, Germany

Abstract. This paper presents a system for another input modality in a multimodal human-machine interaction scenario. In addition to other common input modalities, e.g. speech, we extract head gestures by image interpretation techniques based on machine learning algorithms to have a nonverbal and familiar way of interacting with the system. Our experimental evaluation proofs the capability of the presented approach to work in real-time and reliable.[1]

1 Motivation

Multimodal communication ways are becoming more important for a robust and flexible human-machine interaction in everyday surroundings. Therefore, our objective is to introduce a communication channel providing a natural and intuitive way of simple nonverbal communication. It emulates a common way of showing agreement / disagreement via head gestures, like it is known from human-human dialogs.

Due to its simplicity and omnipresence in every-day life, this contributes to making dialog systems more efficient.

2 Use-Cases

As already mentioned above, head gestures are present in numerous situations. Until now, we have integrated the recognition system in two human-machine interaction scenarios. In the following sections, we will present these two implementations.

[1] All authors contributed equally to the work presented in this paper.

J.A. Jacko (Ed.): Human-Computer Interaction, Part II, HCII 2009, LNCS 5611, pp. 170–177, 2009.

2.1 CoCoMate

Our first area of application is the automated coffee machine "CoCoMate" (Cognitive Coffee Machine) with a pan-tilt camera to facilitate interactions in the visual domain. In Picture 1, the system architecture of our coffee machine is depicted. The user can order various styles of coffee via speech input. This scenario facilitates dialog structures with several user confirmations to questions such as "Would you like to have an espresso?". These confirmations are mostly designed to have a simple yes or no structure. The answers of the user are recognized via speech or head gestures (here: nodding or shaking). Especially in noisy environments (here: grinding coffee beans) the state-of-the-art speech recognition is currently not robust enough to extract the uttered confirmation correctly. Therefore, the additional gesture communication channel can be exploited as complementary information source to deliver the desired semantic information of the user.

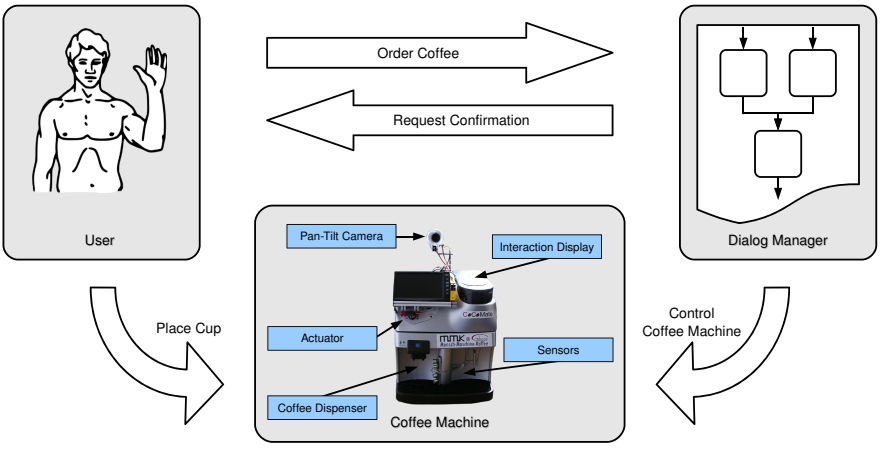

Fig. 1. Controlling CoCoMate: User interacts via a dialog system, which controls the ordering process and the actuators of the machine

2.2 Service Robot

The second implemented use-case is a more wide-spread application scenario concerning human-machine interaction situated in a household (see [1]). A service robot (for more details and examples please refer to [2]) is assisting the human in his daily tasks, e.g. bringing dishes, setting the table, etc. Therefore, the human has to communicate with the robot. To avoid unintended robot actions, it is very important for the human to know, if the robot got the right command. For accomplishing this degree of security, the dialog structure is designed in the following manner, that every action has to be confirmed by the

user. This confirmation can be retrieved via an audio-based command, in this case a commercial speech recognizer software is applied. Additionally, the head of the service robot [3] is equipped with two firewire cameras following the user. One of these cameras is used within the head gesture detecting process.

3 System Overview

The system can be easily adapted to different human-machine dialogs, because the hard- and software requirements are very low. For the integration of the results obtained by the head gesture analysis, many interface modules are imaginable.

3.1 Interfaces

Many different kinds of interfaces are thinkable to constitute the interface between the head gesture software module and the dialog manager. Due to several gained experiences from other projects, we have decided to implement two different interface approaches:

Internet Communication Engine - ICE
ICE is a middleware (similar to SOAP) that allows easy connection of modules by defining interfaces in a meta language. The definitions written in this meta language can be compiled into templates in different programming languages ranging from C++, PHP to Java. Furthermore the modules based on these interfaces can run distributed on different operating systems. In our case, the interface consists of `initialize`, `start`, `stop` and `getResults` methods to control the head gesture recognition process and retrieve the desired information.

Real-time Database
The Real-time Database (RTDB) presented in [4,5,6] is capable to cope with enormous data from varying sensor inputs. Not only the handling of this large amount of data is accomplished easily by the RTDB, but also the processing of these data is done in real-time with a low processing overhead. The original area of application of the Real-time Database was situated in cognitive autonomous vehicles, where the database manages all sensor inputs to keep the vehicle on track.

The RTDB processes objects that can be created and updated by input modules, also called *writers*. In addition to the actual input data, these *writers* have to submit information about the time the data was acquired. Thereby, it is possible for the RTDB to synchronize the data coming in asynchronously from multiple sources and at different sampling rates. Output modules (called *readers*) wait for new objects to process them. In the here presented system, a simple camera-writer and reader is used for the head gesture recognition tool.

As a further advantage of the RTDB, the data stream of the camera-writer can not only be used for on-line head gesture recognition, but can also be recorded to deliver training input for the classification modules.

3.2 Dialog Manager

The controlling unit of the dialog (e.g. when the user is ordering a coffee) is based on a finite state machine (FSM). This dialog manager is responsible for acquiring high-level information and its semantic interpretation gained from sensor modules performing low-level operations. For example: When a human response via head gestures is expected, the dialog manager adjusts the camera on the face of the communication partner under consideration. High-level operations executed by the dialog manager comprise the switching between the states, depending on the currently available sensor inputs.

3.3 Head Gesture Analysis

Referring to the survey presented in [7], this task can be split into three subtasks. First, the location of the face is estimated in the image and a model fitting algorithm provides a face model parameterization. Then, features are extracted that describe the image content. Finally, machine learning techniques derive high-level information.

Face Model fitting
First, the face of the human interaction partner has to be located and extracted from the video data delivered by the RTDB. Therefore, a face detector basing on the well-known Viola Jones approach [8] is utilized to obtain possible locations of human faces. These hypothesis are validated using a model based on skin color [9,10]. In case that more than one validated human face is present in the scene, our first approach is to select the face, which surface surpasses all others in size. Having these information at hand, it is now possible to adjust the pan-tilt camera to focus the user's face in the center of the image plane immediately. As a next step, the generic model of a humanoid face has to be fitted to the one of the current dialog partner.

We integrated an implementation of the algorithm of Viola et al. [11]. A 3D-model forms an abstractions of the visible person's real world face and describes its properties by a parameter vector. In order to extract high-level information, model parameters have to be estimated, describing the face within a given image best. Model fitting solves this task and is often addressed by minimizing an objective function that evaluates, how well a model parameterization fits the processed image.

In contrast to other approaches, the objective function of our method is learned rather than being manually designed. This approach is constituted on general features of ideal objective functions [12]. The main idea behind this applied method is that if the function used to generate training data is ideal, the function learned from the data will also be approximately ideal. Additionally, we provide the learning algorithm with a large number of image features, which characterize the image content. In contrast to humans, the machine learning algorithm is capable to consider this vast amount of data and select the most descriptive features to build the calculation rules of the objective function. This provides both, good runtime performance and high precision.

Feature Extraction
In the second step, descriptive features are extracted, that characterize the displayed head gesture. This step is accomplished by extracting pose information from the fitted rigid 3D face model.

The small amount of model parameters provides a short computational time resulting in real-time capability. Five model parameters – in-plane transition of the face and the three rotation angles – are considered to train a classifier for the recognition of head gestures. Yet, instead of using the absolute parameter values of single images, we calculate the transitional and rotational speed of the face from temporal parameter changes.

Classification
In the third step, the head gesture, performed by the person visible, is obtained from the extracted features.

The feature values gained from the model-fitting process are now analyzed with a continuous Hidden Markov Model (HMM). The HMM was trained from the image data of fourteen different persons performing three basic types of head gestures (nodding, shaking and neutral). Two sequences per test person and head gesture have been captured.

Continuous Hidden Markov Models serve this purpose well for several reasons. First, they inherently consider temporal dynamics of the training data and the test data. Since head gestures are also inherently dynamic, HMMs model their characteristics well and therefore provide profound classification results. Second, HMMs compensate changes in the speed of the performed head gesture during recognition. Therefore, head gestures that are performed slower or faster than depicted in the training data are also recognized. Both advantages allow the HMM to consider temporal aspects, although the data is presented image vice. Therefore, the temporal aspects do not have to be modeled by the designer, for instance via sliding windows.

4 Evaluation

Finally, an experimental evaluation of the recognition system was conducted. This evaluation has the classification accuracy of the Hidden Markov Model under investigation. The reliability of the gained results of the HMM has a crucial impact on the entire human-machine-communication process. This is especially entailed in the reactions of the human user towards the machine, how naturally the communication is perceived and how comfortable he is feeling about using the presented dialog system. Therefore, a six fold cross validation has been conducted to prove the robustness and reliability as well, see Table 1.

For this purpose, we divided the recorded data into six disjunct sets. One of these sets was taken to constitute the testing set and the five remaining sets were used to build the HMM-model. This procedure was conducted six times and the arithmetic mean value of the obtained results can be seen in Table 1. The number of states J of the Hidden Markov Model ranges from four to seven.

Table 1. This table presents recognition rates of our HMM trained with four to seven states. The results are obtained from a 6-fold cross validation.

	Sequence Label		
Classified As	Shaking	Neutral	Nodding
Shaking	**94%**	5%	0%
Neutral	16%	**77%**	7%
Nodding	0%	22%	**78%**
Mean error rate	17.00%		

a) classification result with four states

	Sequence Label		
Classified As	Shaking	Neutral	Nodding
Shaking	**100%**	0%	0%
Neutral	11%	**89%**	0%
Nodding	0%	6%	**94%**
Mean error rate	5.67%		

b) classification result with five states

	Sequence Label		
Classified As	Shaking	Neutral	Nodding
Shaking	**100%**	0%	0%
Neutral	6%	**88%**	6%
Nodding	0%	6%	**94%**
Mean error rate	6.00%		

c) classification result with six states

	Sequence Label		
Classified As	Shaking	Neutral	Nodding
Shaking	**100%**	0%	0%
Neutral	6%	**94%**	0%
Nodding	6%	6%	**88%**
Mean error rate	6.00%		

d) classification result with seven states

Again, the above described six fold cross validation was also applied to determine the recognition accuracy. The best results are achieved for $J = 5$, which can be seen in Table 1, showing the mean error rate and additionally an overview of the accuracy values.

5 Conclusion and Future Work

In this paper we introduce head gesture recognition as a simple and robust communication channel for a human-machine interaction. Due to the following three reasons, the results of the implemented system show high potential of this interaction method. First, this method is similar to the human-human interaction of showing agreement and disagreement (head nodding and head shaking). Second, the system is capable to extract head gestures in real-time with a low resource usage. Third, the learned objective function as well as the gained HMM can be replaced with a more sophisticated version without necessitating further modifications on the head gesture recognizer module.

Future work will lay stress on extending the area of robust applications towards real-life scenarios (lighting conditions, multiple points of view, etc.). Additionally, it is foreseen to integrate facial expressions into the classification process.

Acknowledgment

This ongoing work is supported by the DFG excellence initiative research cluster *Cognition for Technical Systems – CoTeSys*, see `www.cotesys.org` for further details and information. At first, we want to mention here Stefan Sosnowski, the constructor of the EDDIE head [3]. He was a great support for the realization of this project, and he always helped us with words and deeds. The authors further acknowledge the great support of Matthias Göbl for his explanations and granting access to the RTDB repository. In addition, we want to thank all our partners within our CoTeSys-Projects for the fruitful discussions and implementation work to make our visions and ideas become reality.

References

1. Beetz, M., Stulp, F., Radig, B., Bandouch, J., Blodow, N., Dolha, M., Fedrizzi, A., Jain, D., Klank, U., Kresse, I., Maldonado, A., Marton, Z., Mösenlechner, L., Ruiz, F., Rusu, R.B., Tenorth, M.: The assistive kitchen — a demonstration scenario for cognitive technical systems. In: IEEE 17th International Symposium on Robot and Human Interactive Communication (RO-MAN), Muenchen, Germany (2008) (Invited paper)
2. Homepage of Institute of Automatic Control Engineering (LSR), Technische Universität München, Munich, `http://www.lsr.ei.tum.de/research/research-areas/robotics/murola-the-multi-robot-lab`

3. Sosnowski, S., Kuhnlenz, K., Buss, M.: EDDIE - An Emotion-Display with Dynamic Intuitive Expressions. In: The 15th IEEE International Symposium on Robot and Human Interactive Communication. ROMAN 2006, University of Hertfordshire, Hatfield, United Kingdom, September 6-8, 2006, pp. 569–574 (2006)
4. Goebl, M., Färber, G.: A real-time-capable hard- and software architecture for joint image and knowledge processing in cognitive automobiles. In: Intelligent Vehicles Symposium, pp. 737–740 (June 2007)
5. Stiller, C., Färber, G., Kammel, S.: Cooperative cognitive automobiles. In: Intelligent Vehicles Symposium, pp. 215–220. IEEE, Los Alamitos (2007)
6. : Thuy, M., Göbl, M., Rattei, F., Althoff, M., Obermeier, F., Hawe, S., Nagel, R., Kraus, S., Wang, C., Hecker, F., Russ, M., Schweitzer, M., León, F.P., Diepold, K., Eberspächer, J., Heißing, B., Wünsche, H.J.: Kognitive automobile - neue konzepte und ideen des sonderforschungsbereiches/tr-28. In: Aktive Sicherheit durch Fahrerassistenz, Garching bei München, April 7-8 (2008)
7. Pantic, M., Rothkrantz, L.J.M.: Automatic analysis of facial expressions: The state of the art. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(12), 1424–1445 (2000)
8. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Conference on Computer Vision and Pattern Recognition1, pp. 511–518 (2001)
9. Soriano, M., Huovinen, S., Martinkauppi, B., Laaksonen, M.: Skin Detection in Video under Changing Illumination Conditions. In: Proc. 15th International Conference on Pattern Recognition, Barcelona, Spain, pp. 839–842 (2000)
10. Vezhnevets, V., Sazonov, V., Andreeva, A.: A survey on pixel-based skin color detection techniques. In: Proc. Graphicon 2003, pp. 85–92 (2003)
11. Viola, P., Jones, M.J.: Robust real-time face detection. International Journal of Computer Vision (2004)
12. Wimmer, M., Stulp, F., Pietzsch, S., Radig, B.: Learning local objective functions for robust face model fitting. IEEE (PAMI) 30(8) (2008)