Real-Time Face and Gesture Analysis for Human-Robot Interaction

Frank Wallhoff[a], Tobias Rehrl[a]
Christoph Mayer[b], Bernd Radig[b]

[a]Human-Machine Communication
Department of Electrical Engineering and Information Technologies
Technische Universität München
Munich, Germany
[b]Chair for Image Understanding and Knowledge-Based Systems
Computer Science Department
Technische Universität München
Munich, Germany

ABSTRACT

Human communication relies on a large number of different communication mechanisms like spoken language, facial expressions, or gestures. Facial expressions and gestures are one of the main nonverbal communication mechanisms and pass large amounts of information between human dialog partners. Therefore, to allow for intuitive human-machine interaction, a real-time capable processing and recognition of facial expressions, hand and head gestures are of great importance. We present a system that is tackling these challenges. The input features for the dynamic head gestures and facial expressions are obtained from a sophisticated three-dimensional model, which is fitted to the user in a real-time capable manner. Applying this model different kinds of information are extracted from the image data and afterwards handed over to a real-time capable data-transferring framework, the so-called Real-Time DataBase (RTDB). In addition to the head- and facial-related features, also low-level image features regarding the human hand - optical flow, Hu-moments - are stored into the RTDB for the evaluation process of hand gestures. In general, the input of a single camera is sufficient for the parallel evaluation of the different gestures and facial expressions. The real-time capable recognition of the dynamic hand and head gestures are performed via different Hidden Markov Models, which have proven to be a quick and real-time capable classification method. On the other hand, for the facial expressions classical decision trees or more sophisticated support vector machines are used for the classification process. These obtained results of the classification processes are again handed over to the RTDB, where other processes (like a Dialog Management Unit) can easily access them without any blocking effects. In addition, an adjustable amount of history can be stored by the RTDB buffer unit.

Keywords: real-time image processing, gesture recognition, facial expressions

1. INTRODUCTION

The excellence research cluster *Cognition for Technical Systems CoTeSys* attempts to equip technical systems with a high degree of cognition,[1,2] thus facilitating more intelligent and useful reactions of these technical systems towards its surroundings, in particular in human-machine interaction scenarios. The cluster of excellence focuses on human-machine interaction scenarios in two main research fields: ambient living[3] and advanced robotics,[4] especially in an industrial context.[5] In general, for enabling a technical system to provide a cognitive behavior, this system has to be capable of perceiving its human interaction partner in a holistic way. Due to the fact, that humans do not only rely on one interaction method, e.g. speech, a cognitive system should be able to apprehend the human in a multimodal manner.[6] Therefore, we set up a real-time capable ges ture interaction interface – evaluating head and hand gestures as well as facial expressions – for human-robot interaction applicable for different scenarios.

The rest of this paper is organized in the following manner: A brief overview of gesture recognition as well as facial expression analysis in the literature is presented in Section 2. In Section 3, we introduce the demonstration

scenarios for our real-time capable gesture interface, that can be situated either in an assistive household scenario or in an industrial context. Afterwards, in Section 4, we have a closer look on the foundation constituting our real-time capable framework for the gesture interaction analysis. Section 5 considers the features analyzed for the gesture and facial expression classification processes delineated in Section 6. In Section 7 the real-time capable gesture processing is explained. The paper closes with a summary and an outlook over the next planned working steps.

2. RELATED WORK

For showing agreement or disagreement, head gestures are a pleasant way of communication to humans.[7] Nonetheless, in[8] head gestures were applied for controlling operations, like for document browsing. In contrast to head gestures, hand gestures offer a large variety of expression possibilities ranging form pointing gestures over gesticulation towards language-like gestures up to sign language, thus they are very widespread in human-human communications. A classical paper utilizing hand gesture recognition[9] applies Hidden Markov Models[10] to classify the different gestures. Our areas of application are situated in either a typical living environment or in an industrial context, 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.[11] In[12] a remo te control is introduced that utilizes ten predefined hand gestural commands to control a device selected via a pointing gesture. In[13] a system is presented that is capable to recognize hand gestures for human-robot interaction context.

Ekman and Friesen find six universal facial expressions that are expressed and interpreted independent of the cultural background, age or country of origin all over the world.[14] The Facial Action Coding System (FACS) precisely describes the muscle activity within a human face that appear during the display of facial expressions.[15] Some approaches infer the facial expression from rules stated by Ekman and Friesen.[16] Kotsia et al. take this approach by fitting a face model to example images showing either a neutral face or a strongly displayed facial expression and extracting the model parameters.[17] The accuracy of model fitting is of great importance to model-based approaches since the accuracy of facial expression recognition depends directly on the model's capability to reduce the image data and reflect its content. Pure image data without prior data reduction is also applied for facial expression classification. Littlewort et al take this approach.[18] They utilize AdaBoost and Support Vector Machines to determine the facial expression from convolutions of image data with Gabor energy filters and their approach obtains very high recognition results. Local binary patterns are applied in recent work by Shan et al.[19] These approaches suffer from large changes in the face appearance of the face which is usually not provided by standard databases.

Human-robot interaction arises more and more in current research activities, whereas the focus can be laid on several topics. For the cooperation between human and robots in an industrial context, the hybrid assembly constitutes the cooperation form requiring multimodal interaction methods (speech, gestures) for realizing a natural human-machine interaction. A precise overview of human-robot collaboration in assembly lines is presented in.[20] In[21] three robot assistants are introduced for manufacturing and automation. In addition to hybrid assembly, ambient assistive living is a research area, where the integration of a robotic platform in the interaction with humans obtains more importance. In[22] a mobile robotic research platform is presented, which should assist human in their daily life.

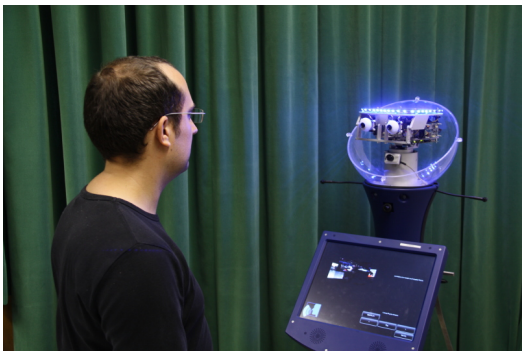
3. SCENARIO

As above-mentioned, the presented system has a generic approach and can therefore be applied to several areas. Up to now, there are two scenarios, where the real-time capable framework can support the human-robot interaction: an assistive household, and industrial hybrid assembly.

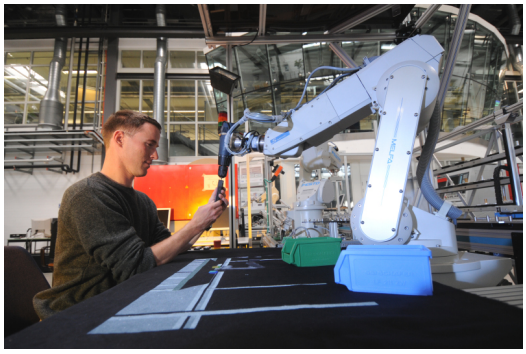
As known from human-human interactions, there are many possible ways that humans utilize for their communication with each other. To introduce this natural and pleasant way of interaction also in our human-robot interaction scenarios, we wanted to extend the speech-based robot control with gestural interactions.

The current starting point for the assistive household scenario is a serving operation. The robot asks the users what they want to drink. Afterwards, the robot brings the ordered drinks to the users. However, this serving operation is only the beginning for setting up an assistive household environment.

In contrast, for the hybrid assembly, there is an demonstration scenario set up, where the basic concepts and framework foundations are suitable for incorporating the gesture as well as the facial expression recognition in the dialog system. The objective for the integration of the gesture recognition in the industrial context is mainly due to the fact that a reliable speech recognition is not always available, since the ambient noise can disturb the speech recognition process.



(a) Assitive household.



(b) Hybrid assembly.

Figure 1.

4. FRAMEWORK

In the subsequent lines, we will have a closer look on the underlying framework facilitating the real-time capable recognition of gestures (head, hand) as well as facial expressions. Despite the fact that the Real-time Database (RTDB) was developed for operations in so-called cognitive autonomous vehicles, the presented framework is also applicable and appropriable for our human-robot interaction scenarios. As it can been seen in,[23,24] the RTDB is capable of dealing with large amount of data input streams, having diverse features (i.e. data rate, packet losses, etc.).

Although, the RTDB is named database, it is more like a sensory buffer that records a certain time period and makes these data available for different modules. This sensory buffer acts as a shared memory, which allows different software-modules the parallel access to the same input data without any blocking effects. Besides a low computational processing overhead of the RTDB, different modules can process the same data originating from one input source. For example different modules can process an image retrieved from one camera, that is the reason why we only need the input of one camera incorporated to classify head and hand gestures as well as facial expressions.

4.1 The Concept of the "RTDB-Writer-Module"

The so-called "RTDB-Writer-Module" is necessary for making processed data available in the RTDB memory. This module constitutes a frame around a different software-module processing varying kinds of data (i.e. audio streams, video data, etc.). In the frame of the "RTDB-Writer-Module" timestamps of the processed and committed data is added to the input data. Due to the fact that the RTDB can process input streams from different data sources, the memory allocation has to be performed according to the requirements of the input data. Besides, the required process handling for controlling and distinguishing different input data streams is accomplished by the RTDB framework.

4.2 The Concept of the "RTDB-Reader-Module"

As counterpart to the "RTDB-Writer-Module" the so-called "RTDB-Reader-Module" constitutes the necessary connection process handling, that a software-module can retrieve data from the RTDB memory. Due to the fact that the RTDB is based on shared memory, different reader-modules can process the same input data stream. Additionally, the access to the stored data can be coordinated and accomplished via the timestamps labeled onto the input data.

4.3 Software Modules for Data Input Processing

As above-mentioned in the scenario description, we located our human-robot interaction in either an ambient living setting or in an industrial context. The human-robot interaction system is capable of processing multimodal input data streams (i.e. different webcams, audio data). Therefore, different kind of "RTDB-Writer-Modules" were implemented. For the processing of the gestures and the facial expression only a "RDTB-Video-Writer" is necessary to retrieve image-data from a webcam. The "RDTB-Video-Writer" delivers raw RGB data from different camera types (i.e. USB webcam, firewire cams) in a common image representation basing on the OpenCV format.[25] Due to the fact that many standard and high-level computer vision algorithms are implemented in the OpenCV library, we choose OpenCV for the image-processing steps.

5. FEATURES

For ensuring a real-time capable behavior of the recognition framework, features are selected, that can be extracted from the image data fast enough to meet real-time requirements. As starting point we apply a three-dimensional model for estimating the features for the facial expression analysis as well as the head gesture classification. The hand gestures are determined in a defined area around the head by setting up region of gesture action ("roga").

5.1 Model Fitting

Model fitting is applied for the analysis of face images to reduce the large amount of image data to a small number of descriptive model parameters by exploiting a priori knowledge about human faces. Finding an accurate model parameterization that matches the image content is a computational challenge, thus model fitting utilizes fast and robust learned displacement experts. The *displacement expert* $f(I, \boldsymbol{p})$ yields a comparable value that suggests an update $\Delta\boldsymbol{p}$ to a parameterized model $\boldsymbol{p}$ in order to better fit an image I.

5.1.1 Facial Expressions

Since facial expression recognition requires accurate information about the position and shape of facial features, we integrate the Candide-III model[26] which has been tailored to face model fitting. An initial estimate of the position of the face is determined with the approach proposed by Viola and Jones,[27] which is implemented in several computer vision libraries. Head rotation or translation do not give any evidence about facial expressions and therefore the only features that are relevant to facial expression classification are the shape parameters. To acquire more robust features, we compute the difference between the parameters of the current image and those of a neutral reference image of the person. The current parameters and the difference between the current parameters and the neutral parameters of the person are stored into a vector ρ_t.

at the cost of classification robustness, compared to more advanced classification methods. The second classifier is a Support Vector Machine SVM.[28] SVMs determine the maximum-margin separating hyperplane to separate binary labeled data. Non-linear hyperplanes are achieved by mapping the data into a higher-dimensional feature space using a kernel. Multi-class SVMs are constructed by splitting the problem into multiple binary classifications. For this paper, we employ a one-versus-one version of the multi-class SVM with a polynomial kernel, where a classifier is trained for every pair of labels separately.

5.1.2 Head Gestures

Applying the model fitting process and the Candide-III face model,[26] we obtain an abstraction of the human face, which yields to a characterizing parameter vector $\boldsymbol{p}$. The three dimensional temporal variation of the pose information of the head is considered for the head gesture classification. Thus in total, six model parameters are evaluated for the recognition of head gestures, which are comprised in the new $\Theta_t = (\Delta px_t, \Delta py_t, \Delta pz_t, \Delta\alpha_t, \Delta\beta_t, \Delta\gamma_t)^T$, where the temporal changes of the face in-plane transition are given by:

$$\begin{aligned}
\Delta px_t &= px_t - px_{t-1}, & \forall t \subseteq \{1, ..., n\}, px_{t=0} &= 0 \\
\Delta py_t &= py_t - py_{t-1}, & \forall t \subseteq \{1, ..., n\}, py_{t=0} &= 0 \\
\Delta pz_t &= pz_t - pz_{t-1}, & \forall t \subseteq \{1, ..., n\}, pz_{t=0} &= 0,
\end{aligned}$$

and the temporal changes of the three rotation angles are given by:

$$\begin{aligned} \Delta\alpha_t &= \alpha_t - \alpha_{t-1} & \forall t \subseteq \{1,...,n\}, \alpha_{t=0} = 0 \\ \Delta\beta_t &= \beta_t - \beta_{t-1} & \forall t \subseteq \{1,...,n\}, \beta_{t=0} = 0 \\ \Delta\gamma_t &= \gamma_t - \gamma_{t-1} & \forall t \subseteq \{1,...,n\}, \gamma_{t=0} = 0. \end{aligned}$$

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

5.2 Hand Gestures

Similar to the approach mentioned in,[29] 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. To further constrain the hand gesture recognition, a region of gesture action ("roga") is set up. The area for the hand gesture interaction is located with regard to the position of the head. In the subsequent processing step, the "roga" is compared with above gained information of possible face candidates to eliminate false hand candidates. Then, the obtained processed image delivers the input for the hand gesture classification. From the obtained hand contour image $Im(x, y, t)$ for time instance t, the center of gravity (cx_t, cy_t) is determined by utilizing the moments.

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

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

5.2.1 Position Related Features

The center of gravity constitutes the foundation for the position-related features. There are two features covering the temporal change Δt of the center of gravity.

$$\begin{aligned} \Delta cx_t &= cx_t - cx_{t-1}, & \forall t \subseteq \{1,...,n\}, cx_{t=0} = 0 \\ \Delta cy_t &= cy_t - cy_{t-1}, & \forall t \subseteq \{1,...,n\}, cy_{t=0} = 0 \end{aligned}$$

5.2.2 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 so-called Hu moments[30] 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. The seven input features are shown below.

$$
\begin{aligned}
\eta_{ij} &= \frac{\mu_{ij}}{\mu_{00}^{\left(1+\frac{i+j}{2}\right)}} \\
hu_1 &= \eta_{20} + \eta_{02} \\
hu_2 &= (\eta_{20} - \eta_{02})^2 + (2\eta_{11})^2 \\
hu_3 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
hu_4 &= (\eta_{30} + \eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\
hu_5 &= (\eta_{30} - 3\eta_{12})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\
&\quad + (3\eta_{21} - \eta_{03})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\
hu_6 &= (\eta_{20} - \eta_{02})[(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2] \\
&\quad + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03}) \\
hu_7 &= (3\eta_{21} - \eta_{03})(\eta_{30} + \eta_{12})[(\eta_{30} + \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] \\
&\quad + (\eta_{30} - 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - (\eta_{21} + \eta_{03})^2]
\end{aligned}
$$

The vector $\xi_t = (hu_{1_t}, hu_{2_t}, hu_{3_t}, hu_{4_t}, hu_{5_t}, hu_{6_t}, hu_{7_t})^T$ comprises the shape related features, whereas the vector $\Lambda_t = (\Delta cx_t, \Delta cy_t)^T$ comprises the position 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.

6. CLASSIFICATION

In this section we present our applied classification methods: For the gesture recognition Hidden-Markov-Models (HMMs) are utilized, whereas for the classification of the facial expressions both decision tree and sophisticated support-vector-machines (SVMs) are applied.

6.1 Facial Expression Classification

Two databases serve as training and testing data for the facial expression classification: The Cohn-Kanade Facial Expression Database (CK) and the MMI Face-Database. Image sequences of the Cohn-Kanade database start with a neutral image and depict the evolving facial expression until the apex is reached in the final image. In addition, the MMI data contains also images of decreasing facial expressions. To determine the point of transition between a neutral face and a depicted facial expression, the expression intensity is modeled to increase linearly between neutral images and apex frames. All images with a intensity of less than 0.3 are considered neutral, the remainder to depict facial expression. The model parameters of single images are exploited to recognize facial expressions. We evaluate the classification algorithm with different sets of features. The classifier C1 is trained on model parameters of a single image only and therefore it is applicable to single images, without requiring a comparative image depicting a neutral facial configuration. Classifier C2 is trained on model parameter differences between the image of interest and the first image of a sequence, which serves as neutral reference image. Classifier C3 is trained on the combination of both features. The classifiers are evaluated by a 10-fold cross-validation on both databases as presented in Table 1. The first classification algorithm used a *decision tree algorithm*,[31] whereas the second classifier applied *Support Vector Machine* (SVM).[28] Decision trees are tree structures that branch on single feature comparisons on inner nodes and return a single class in their leaf nodes. They are implemented due to their fast execution which contributes to the real-time capability of the complete system. Support Vector Machines fit a maximum-margin hyper plane in the training data after transforming the original data to a higher dimensional space using a kernel function. A standard radial basis function kernel is implemented in our approach. The obtained results from the both approaches can be seen in Table 1.

	CK_{C1}	MMI_{C1}	CK_{C2}	MMI_{C2}	CK_{C3}	MMI_{C3}
decision tree	83.6%	92.6%	84.7%	91.8%	86.1%	92.0%
SVM	87.9%	93.1%	88.4%	92.7%	91.4%	87.8%

Table 1. Recognition accuracies for different feature sets and classifiers. The numbers are given for Cohn-Kanade Facial Expression Database and MMIFace Database, respectively.

6.2 Gesture Classification

Continuous Hidden Markov Models (HMMs)[10] are applied to classify the hand gestures as well as the head gestures. A Hidden Markov Model $\lambda = (A, b, \pi)$ relies on N internal emitting states x_i, an initial state distribution π_i, a state transition matrix A, and the (continuous) production probability vector $\vec{b} = [b_1 \dots b_J]^{\mathrm{T}}$ to calculate the probability that a sequence of feature vectors is produced by a particular pattern sequence, i.e. a certain gesture.

The state transition matrix A comprises the transition probabilities $a_{x_i x_j}$ from state x_i to state x_j (1st order Markov Model). The elements b_j in a certain state x_j for a D-dimensional observation $\vec{y_j}$ are given by a multivariate Gaussian distribution consisting of a mean value vector $\vec{\mu}_j$ and a covariance matrix Σ_j.

$$b_j(\vec{y_j}, \vec{\mu}_j, \Sigma_j) = \frac{1}{\sqrt{(2\pi)^D |\Sigma_j|}} e^{-\frac{1}{2}(\vec{y_j}-\vec{\mu}_j)^T \Sigma_j^{-1} (\vec{y_j}-\vec{\mu}_j)} \tag{1}$$

They describe the probability of a given observation $\vec{y_j}$ in a state x_j.

During the training phase the unknown parameters in A and $\vec{b}$ are calculated. For this purpose the well-known Baum-Welch-Estimation procedure[10] is applied.

The Hidden Markov Model parameters are computed by the following maximum-likelihood decision:

$$\lambda = \underset{\lambda}{\operatorname{argmax}}\ P(Y|\lambda) \tag{2}$$

Where Y represents a vector of observations and λ a model parameterization.

The trained HHMs 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. The obtained results from the HMM approach for both the head gestures and hand gestures can be seen in Table 2 and in Table 3, respectively.

Head gesture classification result using HMMs			
Classified as	Sequence Label		
	Shaking	Neutral	Nodding
Shaking	95%	5%	0%
Neutral	5%	85%	10%
Nodding	00%	0%	100%

Table 2. This table presents recognition rates of a HMM-based classification for the head gestures. The results are obtained from a five-fold cross validation.

7. GESTURE PROCESSING

In this section, we will present the system processing steps during execution time. Our approach can roughly be subdivided into four major processing operations: data acquisition, preprocessing, feature extraction, and classification. All performed steps are interrelated to the RTDB. Three different types of classification processes can be distinguished: head gestures, hand gestures, and facial expressions. In the Figure 2, the general processing steps are depicted for the example of the head gesture analysis.

Hand gesture classification result using HMMs						
Classified as	Sequence Label					
	Fist right	Fist left	Hand right	Hand left	Hand up	Hand down
Fist right	85%	0%	0%	15%	0%	0%
Fist left	0%	100%	0%	0%	0%	0%
Hand right	5%	0%	95%	0%	0%	0%
Hand left	0%	0%	5%	95%	0%	0%
Hand up	0%	5%	0%	0%	95%	0%
Hand down	0%	0%	0%	0%	0%	100%

Table 3. This table presents recognition rates of a HMM-based classification for the hand gestures. The results are obtained from a five-fold cross validation.

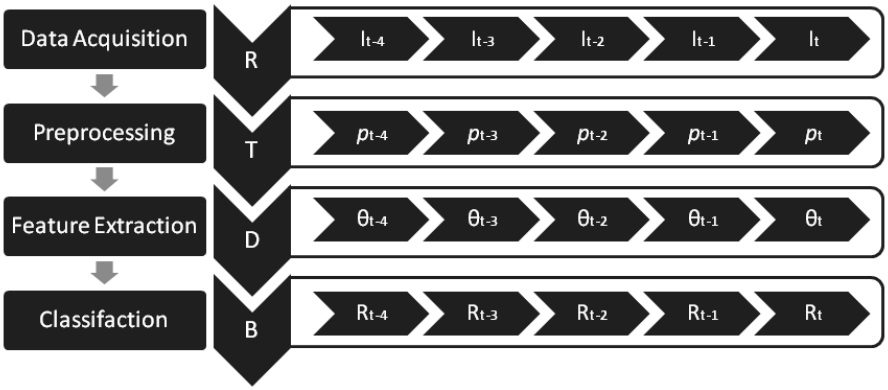


Figure 2. General gesture processing overview. In this case for the head gestures.

7.1 Facial Expression Processing

The raw image data I_t is obtained in the *data acquisition* step from a webcam for the time instance t. From this raw image, the characterizing parameter vector $\boldsymbol{p}_t$ is calculated in the *preprocessing* step. If the obtained image is the first image of a specific person, the parameters are stored as neutral reference parameters $\boldsymbol{p}_0$. Note, that therefore the system requires the first image of a person to be of neutral expression. In the *feature extraction* step, the necessary information for classifying facial expressions are extracted from $\boldsymbol{p}_t$ and stored in the new vector ρ_t. Following the evaluation performed in Section 6.1, the difference between the current shape parameters and the neutral reference shape parameters (feature set C2) are computed. Therefore ρ_t is given by $\rho_t = \boldsymbol{p}_t - \boldsymbol{p}_0$. In the final *classification* step, ρ_t is handed over to the SVM-based classification process yielding a classification result R_t. In Figure 3, the six different facial expressions (anger, disgust, fear, laughing, sadness, surprise) are shown.

Figure 3. Overview of the facial expression: anger, disgust, fear, laughing, sadness, surprise (from left to right).

7.2 Head Gesture Processing

The raw image data I_t is obtained in the *data acquisition* step from a webcam for the time instance t. From tis raw image, the characterizing parameter vector $\boldsymbol{p}_t$ is calculated in the *preprocessing* step. In the *feature extraction* step, the necessary information for classifying head gestures are extracted from $\boldsymbol{p}_t$ and stored in the new vector Θ_t comprising the temporal variations of these information between time step t and $t-1$. In the final *classification* step, a sequence of vectors $\Theta_t \dots \Theta_{t-T_{head}}$ (T_{head} mean time duration of a head gesture) are

handed over to the HMM-based classification process yielding in a classification result R_t. In Figure 4 a nodding sequence is shown from the three possible classes (neutral, nodding, shaking) of the head gestures processing.

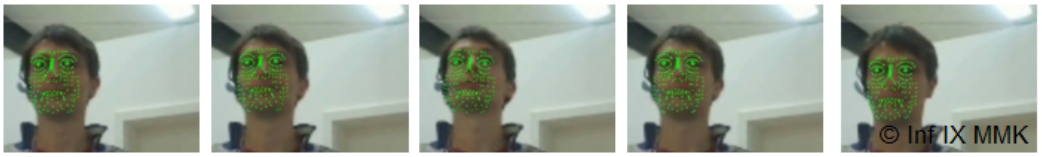


Figure 4. Head gesture processing: a nodding sequence is depicted, where the model is fitted on to the face.

7.3 Hand Gesture Processing

The raw image data I_t is obtained in the *data acquisition* step from a webcam for the time instance t. In the *preprocessing* step, a skin color image within a certain defined region ("roga") is obtained and stored into the image Im_t is calculated. In the *feature extraction* step, the necessary information for classifying hand gestures are extracted from the skin color image Im_t and stored in the vector ζ_t, which comprises both the position related features (Λ_t) and the shape related features (ξ_t). In the final *classification* step, a sequence of vectors $\zeta_t \dots \zeta_{t-T_{hand}}$ (T_{hand} mean time duration of a hand gesture) are handed over to the HMM-based classification process yielding in a classification result R_t. In Figure 5 a hand motion sequence (hand down) is shown from the six possible classes (hand up/down/right/left, fist up/down) of the hand gestures processi ng.

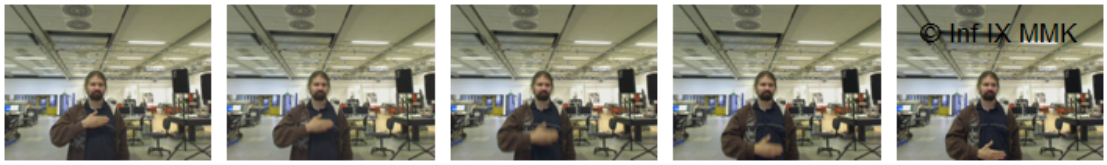


Figure 5. Hand gesture processing: a sequence of a hand down movement is depicted.

8. CONCLUSION

We introduced a real-time capable framework for the recognition of gestures (head, hand) as well as facial expressions. HMMs are applied to classify head gestures as well as hand gestures, whereas the facial expressions are recognized via SVMs. Subsequent steps will tackle the robustness of extracted features by integrating additional depth information from either using time-of-flight cameras or stereo-vision system to improve hand segmentation and gesture classification as well. Finally, a fusion of results originating from different classification methods is imaginable to provide a better and more robust classification result.

ACKNOWLEDGMENTS

This ongoing work is partially supported by the DFG excellence initiative research cluster *Cognition for Technical Systems – CoTeSys*, see `www.cotesys.org` for further details and information. The authors further acknowledge the great support of Matthias Göbl for his explanations and granting access to the RTDB repository.

REFERENCES

[1] Vernon, D., Metta, G., and Sandini, G., "A Survey of Artificial Cognitive Systems: Implications for the Autonomous Development of Mental Capabilities in Computational Agents," *Evolutionary Computation, IEEE Transactions on* **11**(2), 151–180 (2007).

[2] Anderson, M. L., "Embodied cognition: A field guide," *Artificial Intelligence* **149**, 91–130 (September 2003).

[3] 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., and 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.

[4] Lenz, C., Nair, S., Rickert, M., Knoll, A., Rösel, W., Bannat, A., Gast, J., and Wallhoff, F., "Joint Actions for Humans and Industrial Robots: A Hybrid Assembly Concept," in [*Proc. 17th IEEE International Symposium on Robot and Human Interactive Communication*], (2008).

[5] Zäh, M. F., Lau, C., Wiesbeck, M., Ostgathe, M., and Vogl, W., "Towards the Cognitive Factory," in [*Proceedings of the 2nd International Conference on Changeable, Agile, Reconfigurable and Virtual Production (CARV)*], (July 2007).

[6] Jaimes, R. and Sebe, N., "Multimodal human computer interaction: A survey," 15–21 (2005).

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

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

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

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

[11] Kölsch, M. and Turk, M., "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)*], 158 (2004).

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

[13] Hasanuzzaman, M., Zhang, T., Ampornaramveth, V., Gotoda, H., Shirai, Y., and Ueno, H., "Adaptive visual gesture recognition for human-robot interaction using a knowledge-based software platform," *Robot. Auton. Syst.* **55**(8), 643–657 (2007).

[14] Ekman, P., "Universals and cultural differences in facial expressions of emotion," in [*Nebraska Symposium on Motivation 1971*], Cole, J., ed., **19**, 207–283, University of Nebraska Press, Lincoln, NE (1972).

[15] Ekman, P., "Facial expressions," in [*Handbook of Cognition and Emotion*], Dalgleish, T. and Power, M., eds., John Wiley & Sons Ltd, New York (1999).

[16] Ekman, P. and Friesen, W., [*The Facial Action Coding System: A Technique for The Measurement of Facial Movement*], Consulting Psychologists Press, San Francisco (1978).

[17] Kotsia, I. and Pitas, I., "Facial expression recognition in image sequences using geometric deformation features and support vector machines," *IEEE Transactions On Image Processing* **16**(1), 172–187 (2007).

[18] Littlewort, G., Bartlett, M. S., Fasel, I., Susskind, J., and Movellan, J., "Dynamics of facial expression extracted automatically from video," *Image and Vision Computing* **24**, 615–625 (2006).

[19] Shan, C., Gong, S., and McOwan, P. W., "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing* **27**, 803–816 (2009).

[20] Krüger, J., Lien, T., and Verl, A., "Cooperation of human and machines in assembly lines," (2009).

[21] Schraft, R. D., Helms, E., Hans, M., and Thiemermann, S., "Man-machine-interaction and co-operation for mobile and assisting robots."

[22] Graf, B., Parlitz, C., and Hägele, M., "Robotic home assistant care-o-bot®3 product vision and innovation platform," in [*Proceedings of the 13th International Conference on Human-Computer Interaction. Part II*], 312–320, Springer-Verlag, Berlin, Heidelberg (2009).

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

[24] Stiller, C., Färber, G., and Kammel, S., "Cooperative cognitive automobiles," in [*Intelligent Vehicles Symposium, 2007 IEEE*], 215–220 (June 2007).

[25] Bradski, G. and Kaehler, A. in [*Learning OpenCV: Computer Vision with the OpenCV Library*], O'ReillyPress (2008).

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

[27] Viola, P. and Jones, M. J., "Robust real-time face detection," *International Journal of Computer Vision* **57**(2), 137–154 (2004).

[28] Cristianini, N. and Shawe-Taylor, J., [*An Introduction to Support Vector Machines and other kernel-based learning methods*], Cambridge University Press (2000).

[29] Nickel, K. and Stiefelhagen, R., "Visual recognition of pointing gestures for human-robot interaction," *Image and Vision Computing* , 1875–1884, Elsevier (2007).

[30] Hu, M., "Visual pattern recognition by moment invariants," *IRE Transaction on Information Theory* , 179–187 (1963).

[31] Quinlan, J. R., [*C4.5: Programs for Machine Learning*], Morgan Kaufmann, San Mateo, California (1993).