Robust Recognition of Complex Gestures for Natural Human-Robot Interaction

Maren Bennewitz*, Tobias Axenbeck*, Sven Behnke†, and Wolfram Burgard*
*Institute for Computer Science, University of Freiburg, D-79110 Freiburg, Germany
†Institute for Computer Science, University of Bonn, D-53117 Bonn, Germany

***Abstract*— Robots coexisting with humans in everyday environments should be able to interact with them in an intuitive way. This requires that the robots are able to recognize typical gestures performed by humans such as pointing gestures, waving, or head shaking/nodding. We present a system that is able to spot and recognize complex, parameterized gestures from data of a monocular camera. To represent people, we locate their faces and hands using trained classifiers and track them over time. We use few, expressive features extracted from this compact representation as input to hidden Markov models (HMMs). First, we segment the gestures into distinct phases and train HMMs for each phase separately. Then, we construct composed HMMs, which consist of the individual phase-HMMs. Once a specific phase is recognized, we estimate the parameter of a gesture such as the target of a pointing gesture. As we demonstrate in the experiments, our system is able to robustly spot and recognize a variety of complex gestures.**

I. Introduction

Robotic assistants designed to communicate with untrained users must be able to interact with them in a natural way. Our humanoid robot (see Fig. 1) is able to generate a variety of natural arm and head gestures that support its speech [1]. When evaluating questionnaires filled out by people who interacted with the robot at former public demonstrations, we discovered that they were confused by the asymmetry between action generation and perception. The robot's visual perception of people was limited to head position and size at that time. To reduce this asymmetry, it is necessary that the robot also recognizes gestures performed by humans. This requires robust and accurate tracking of human body parts as well as the ability to spot and recognize typical gestures in order to infer non-verbal signals of attention and intention.

We present a system that is able to spot and recognize complex gestures from data of a monocular camera. We consider gestures performed with head and arms, such as head shaking/nodding or hand waving as well as parameterized gestures, such as pointing gestures or gestures indicating the size of objects. Figure 2 shows examples of such typical gestures performed by humans during an interaction.

The contribution of our work is a robust and fast gesture recognition method that relies only on data of a monocular camera (no stereo). In contrast to previous approaches relying on monocular image sequences (e.g., [7, 4]), our system works under realistic settings such as varying and difficult lighting conditions, multiple people, and cluttered background. On a notebook computer, we achieve a frame rate of 20 fps and are able to spot gestures as well as to recognize them, i.e., our system distinguishes between previously learned gestures and irrelevant or unconscious movements.

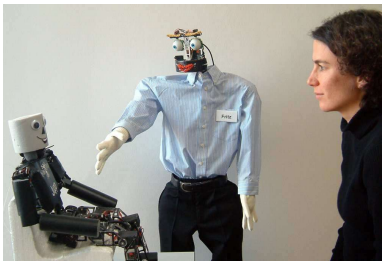

Fig. 1. Our humanoid robot interacts with people using multiple modalities such as speech, facial expressions, eye-gaze, and gestures.

(a) (b) (c)

Fig. 2. Snapshots of typical gestures analyzed in our experiments: (a) waving, (b) indicating the size of an object, and (c) pointing to an object. Our system works robustly even with cluttered background and under different lighting conditions. The bounding boxes highlight detected faces and hands.

Our approach proceeds in three stages. First, we locate faces and hands in the images and update a probabilistic belief which tracks detected faces and hands over time. Second, we extract features from this compact representation of humans. Finally, these features are used as input to Hidden Markov Models (HMMs) which are trained for individual phases of the gestures. Our system recognizes a variety of complex gestures and can estimate their parameters. Existing techniques for parameter estimation of gestures either concentrate on pointing gestures only [3, 5] or rely on the assumption that the whole gesture can be observed [11]. In contrast to that, our approach allows for the estimation of parameters for general gestures once a specific phase is recognized.

II. Representation and Tracking of Humans

For locating faces and hands in the images, we use the object detection framework proposed by Viola and Jones [9] and train reliable and fast classifiers which operate on grey-scale images. To speed-up the search for hands and to increase robustness, we use an adaptive skin color model (which is initially based on the detected face) and constrain the search to skin-colored regions.

We train two kinds of hand classifiers: a *generic* classifier that detects hands and rejects non-hands and a *specific* classifier that is able to discriminate right hands from left ones. Our hand detection system proceeds in two stages. First, the generic hand detector is applied to skin-colored regions. In case it succeeds, the specific hand classifier is applied. In contrast to other approaches [2, 6], our system is able to robustly locate and track hands with a large number of substantially different shapes and to furthermore determine whether a hand is a left or right one.

We maintain a probabilistic belief about the existence of people and the positions of their faces and hands over time. Using this belief, our system improves robustness, can deal with false detections, and is not restricted to a single person.

Additionally, we track the 3D head pose of people. We use an appearance-based approach [8] which locates distinctive facial features. The positions of the features within the face bounding box serve as input to a neural network which computes the three Euler angles of rotation around the neck.

III. Learning and Recognizing Complex Gestures

In our work, we focus on typical gestures performed by humans during an interaction. We currently consider six different types of gestures:

1) *Waving*: One-handed gesture.
2) *Pointing*: Parametric one-handed gesture.
3) *Thisbig*: This parametric two-handed gesture is carried out to indicate the size of an object.
4) *Dunno*: This two-handed gesture is used to express ignorance (informal short for *don't know*).
5) *Head shaking*.
6) *Head nodding*.

A. Gesture Modeling

To model the complex arm gestures *Waving*, *Pointing*, and *Thisbig*, we use three phases: the preparation phase which is an initial movement before the main gesture, the hold phase which characterizes the gesture, and the retraction phase in which the hand moves back to a resting position. Our motivation behind this segmentation is that once the hold phase is recognized, the parameters of *Pointing* and *Thisbig* can be estimated. Furthermore, this segmentation supports the modeling of *Waving* during which similar movements are repeated several times. The less complex gestures *Dunno* and *Head shaking/nodding* are modeled monolithically. We train individual HMMs for each phase of a gesture separately. Accordingly, we train an overall number of 12 HMMs for the gestures/gesture phases.

In our experiments, continuous left-right HMMs with 3-5 (non-skip) states and a mixture of two Gaussians as output distribution performed best to learn the gestures. We use Viterbi training and the Baum-Welch algorithm to estimate for an HMM λ the transition probabilities a_{ij}^{λ} between states i and j and the observation probabilities $b_j^{\lambda}(o)$ for a state j given an observation o.

To be able to identify movements not corresponding to any learned gesture, we train an additional model. Here, we follow the approach presented by Yang *et al.* [12] and build a HMM by copying all states from all trained models and arrange them in a fully connected HMM with smoothed output probabilities.

B. Gesture Recognition via Composed HMMs

The gesture phases appear in a specific order which has to be considered during recognition. Fig. 3 illustrates the HMM topology for one- and two-handed gestures as well as for head gestures. As indicated by the arrow, the hold phase can occur several times or last differently long. To identify the most likely gesture given a composed HMM, we apply the Viterbi algorithm [10]. The Viterbi algorithm computes the state sequence with maximum likelihood given an observation sequence $O_{1:T} = o_1, \ldots, o_T$. For the HMM λ, the likelihood of the best state sequence of length t ending in state j is recursively defined as

$$\delta_t(j) \quad = \quad \max_{1 \leq i \leq N^{\lambda}} \delta_{t-1}(i) a_{ij}^{\lambda} b_j^{\lambda}(o_t), \delta_1(j) = \pi_j^{\lambda} b_j^{\lambda}(o_1). \quad (1)$$

Here, a^{λ} and b^{λ} are the parameters of λ, N^{λ} is the number of states, and π_j^{λ} specifies the initial state distribution. The algorithm terminates with the computation of the most likely path x_T^* (which is found via backtracking) and its probability P^*

$$P^* \quad = \quad \max_{1 \leq i \leq N^{\lambda}} \delta_T(i). \quad (2)$$

In theory, it would be possible to model one- and two-handed gestures in one large HMM. However, to reduce the amount of necessary training data and to improve recognition accuracy, we use separate HMMs for one- and two-handed gestures. Since the HMMs with differently dimensional input features cannot be compared directly, we consider the two-handed HMM if and only if the HMMs for the right and left hand report the same most-likely gesture. This heuristics is applicable since all our two-handed gestures are symmetric.

C. Input Features

As input to the HMMs, we use few, expressive features extracted from the trajectories of head and hands. First, we transform the position of the hands into coordinates relative to the head position and normalize the coordinates with respect to the size of the face bounding box. For one-handed gestures, we use polar coordinates in the image with the head as origin and the velocity. Accordingly, the feature vector $\mathbf{f}_{one}$ is defined as

$$\mathbf{f}_{one} = (r, \phi, v). \quad (3)$$

Here, r is the distance of the hand to the head, ϕ is the angle, and v is the velocity.

Since the two-handed gestures we consider are symmetric, we measure the difference in x/y-direction of their left and right hand coordinates $(x_t^{l/r}, y_t^{l/r})$ at time t in the features $d_x = |x_t^l| - |x_t^r|$ and $d_y = y_t^l - y_t^r$. Furthermore, we record the sum of the y-coordinates of the hands in the

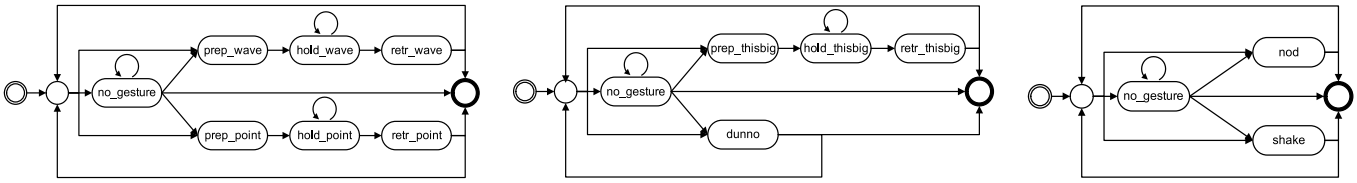

Fig. 3. Composed HMM consisting of the individual phase-HMMs. The first two for one- and two-handed gestures, and the right one for head gestures.

feature $y^{lr} = y_t^l + y_t^r$ and consider the change of the hand coordinates in x-direction

$$\Delta x^l x^r = |x_t^l| - |x_{t-1}^l| + |x_t^r| - |x_{t-1}^r|. \quad (4)$$

As a final feature, we consider the velocities of the hands $v^{lr} = v_t^l + v_t^r$. Thus, the feature vector $\mathbf{f}_{two}$ is defined as

$$\mathbf{f}_{two} = (d_x, d_y, y^{lr}, \Delta x^l x^r, v^{lr}). \quad (5)$$

The head gestures nodding and shaking are described by a feature vector $\mathbf{f}_{head}$ which consists of the three Euler angles of rotation roll, pitch, and yaw as well as their velocities

$$\mathbf{f}_{head} = (\theta^r, \theta^p, \theta^y, v^{\theta_r}, v^{\theta_p}, v^{\theta_y}). \quad (6)$$

D. Estimating Parameters of Gestures

Currently, we consider two parameterized gestures: *Thisbig* and *Pointing*. The corresponding parameters are estimated during the hold phase of the respective gesture. For *Thisbig*, the estimation is done straightforwardly using a learned mapping to estimate the distance of the person to the camera given the bounding box size of the face.

For the estimation of pointing targets, we use of the three rotation angles of the head pose. We assume that people are looking to the object of interest they want to draw the attention to and that the head pose coincides with the gaze direction. Furthermore, we assume the 3D positions of potential pointing targets to be known. First, we estimate the 3D position of the head using the above mentioned mapping from bounding box size to distance. Starting from that position, we construct a straight line in direction of the roll, pitch, and yaw angle of the head pose. Finally, we determine the object which has the closest distance to that line.

IV. Experiments

We performed a series of experiments in order to evaluate our approach. To collect training data, we asked five different people to perform gestures in a distance of 1.5-2.5m to the camera. We chose two different locations, different lighting conditions, and different backgrounds (see Fig. 2). We recorded and processed the videos with a rate of 20fps and used a resolution of 640×480 pixel. We had a database consisting of 75 samples per gesture which we manually labeled, i.e., we marked the start and the end of each gesture as well as the beginning and end of the hold phase.

A. Gesture Recognition

After training the phase-HMMs for the hand gestures, we tested their ability in distinguishing the individual gesture phases (preparation (p), hold (h), and retraction (r) phase). We used the Viterbi path and counted the number of correctly recognized gesture phases from the number of all test sequences. Tab. I shows the percentage of correctly recognized segments for one-handed gestures. As can be seen, using the extracted features, the individual phases of one-handed gestures can correctly be recognized. Only one error occurs for a segment containing a *retr_point* phase which is classified as *retr_wave*. This can be explained by the fact that both retraction phases contain similar movements in the end. When considering a whole observation sequence consisting of all three phases, this error does not occur since the preparation and hold phase are correctly recognized. For the recognition of two-handed gestures shown in Tab. II, it can be seen that in a single test sequence, the phases of *Thisbig* are classified as *Dunno*. When sequences in which persons are not performing any gesture are included into the test set, we achieve an overall recognition rate of 90% for one- as well as for two-handed gestures. The largest part of this error results from the fact that it sometimes happens that *no_gesture* phases are classified as the preparation phase of a gesture.

The following experiment is designed to evaluate the performance of our system on sequences containing whole gestures. We computed the Viterbi path in the composed HMMs at each time step and counted how often the most likely hypothesis corresponds to the true gesture. Fig. 4 shows the results for all six gestures. As can be seen, the gestures can be reliably recognized after processing only few frames. Nodding seems to be most difficult to recognize because sometimes people barely move their head. And, again, we made the observation that *Thisbig* sometimes tends to be classified as *Dunno*.

To better evaluate the ability of our HMMs to distinguish arm gestures, we performed experiments in which we computed for a given observation sequence the Viterbi path and its likelihood for all individual gesture HMMs consisting of the corresponding phase-HMMs (i.e., we did not use the composed HMMs here). We then computed the joint probability $P(g^l, g^r)$ of the gesture g^l of the left and the gesture g^r of the right hand. Fig. 5 plots the evolution of the probabilities of the gestures over time for a sequence in which a person waved with the left hand. In the beginning, the person was not performing any meaningful gesture and, thus, the *no_gesture* model had the highest probability. Afterwards, the probability of the correct gesture increased.

TABLE I

RECOGNITION OF ONE-HANDED GESTURE PHASES.

	p_wave	*h_wave*	*r_wave*	*p_point*	*h_point*	*r_point*	*rec. rate*
p_wave	25	0	0	0	0	0	100%
h_wave	0	25	0	0	0	0	100%
r_wave	0	0	25	0	0	0	100%
p_point	0	0	0	25	0	0	100%
h_point	0	0	0	0	25	0	100%
r_point	0	0	1	0	0	24	96%

TABLE II

RECOGNITION OF TWO-HANDED GESTURE PHASES.

	dunno	*p_thisbig*	*h_thisbig*	*r_thisbig*	*rec. rate*
dunno	25	0	0	0	100%
p_thisbig	1	24	0	0	96%
h_thisbig	1	0	24	0	96%
r_thisbig	1	0	0	24	96%

B. Parameter Estimation

Finally, we asked people to point to predefined targets. We positioned eight different targets within a range of 1.5m to the camera and at different heights. The hold phase of all 66 pointing gestures was identified and the correct target was estimated in 80% of all cases.

Second, we asked people to indicate the size of objects. We told them to indicate the sizes 25cm, 50cm, 100cm, and 150cm and estimated the parameter in the hold phase. We performed 32 experiments and counted the nearest neighbor class of each estimate. Our system was able to determine the correct class in 94% of all cases.

C. Videos

Illustrating videos can be found at our web page[1]. The videos show the robustness of our approach to recognize complex gestures performed by different people. As the experiments demonstrate, gestures can reliably be recognized even under varying lighting conditions and with cluttered background.

V. CONCLUSIONS

We presented an approach to robustly recognize gestures from data of a monocular camera. We consider typical gestures performed by humans during an interaction such as nodding or pointing. To represent people, we locate and track their heads and hands. We use few, expressive features extracted from this compact representation as input to HMMs. We segment complex gestures into three phases and train HMMs for each phase separately. We then construct HMMs composed of the individual phase-HMMs. Using the distinction between different phases, we are able to estimate parameters of gestures as soon as a certain phase is recognized.

Our approach has been implemented and evaluated on a humanoid robot. As the experimental results show, our system is able to reliably spot and recognize gestures, i.e., it distinguishes between previously learned gestures and irrelevant or unconscious movements.

[1] http://www.informatik.uni-freiburg.de/~maren/animations-gestures.html

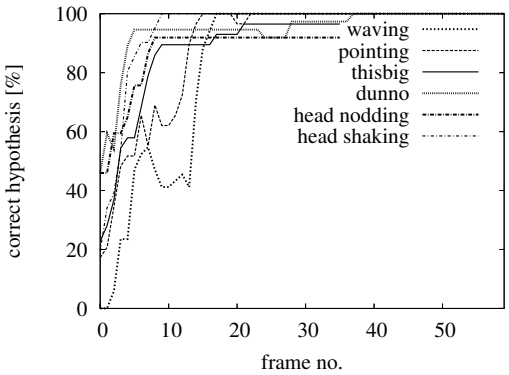

Fig. 4. Number of frames after which the most likely hypothesis is the correct gesture.

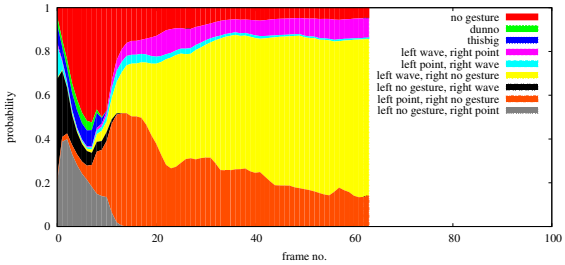

Fig. 5. Evolution of the probabilities of the gestures over time for an experiment in which a person waved with the left hand.

ACKNOWLEDGMENT

This project is supported by the DFG, grant BE 2556/2-2 and by the BMBF project DESIRE.

REFERENCES

[1] M. Bennewitz, F. Faber, D. Joho, and S. Behnke. Fritz – A humanoid communication robot. In *Proc. of the IEEE Int. Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2007.

[2] M. Kolsch and M. Turk. Robust hand detection. In *Proc. of the Sixth IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*, 2004.

[3] C. Martin, F.-F. Steege, and H.-M. Gross. Estimation of pointing poses for visual instructing mobile robots under real-world conditions. In *Proc. of the 3rd European Conference on Mobile Robots (ECMR)*, 2007.

[4] J. A. Montero and L. E. Sucar. Feature selection for visual gesture recognition using hidden Markov models. In *Proc. of the Fifth Mexican Int. Conf. in Computer Science*, 2004.

[5] K. Nickel and R. Stiefelhagen. Detection and tracking of 3D-pointing gestures for human-robot-interaction. In *Proc. of the IEEE/RSJ Int. Conf. on Humanoid Robots (Humanoids)*, 2004.

[6] E.-J. Ong and R. Bowden. A boosted classifier tree for hand shape detection. In *Proc. of the Sixth IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*, 2004.

[7] G. Rigoll, A. Kosmala, and S. Eickeler. High performance real-time gesture recognition using hidden Markov models. In *Proc. of the Int. Gesture Workshop on Gesture and Sign Language in Human-Computer Interaction*, 1998.

[8] T. Vatahska, M. Bennewitz, and S. Behnke. Feature-based head pose estimation from images. In *Proc. of the IEEE/RSJ Int. Conf. on Humanoid Robots (Humanoids)*, 2007.

[9] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. of the IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2001.

[10] A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. on Information Theory*, 13(2):260–269, 1967.

[11] A.D. Wilson and A.F. Bobick. Parametric hidden Markov models for gesture recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(9):884–900, 1999.

[12] H.-D. Yang, A.-Y. Park, and S.-W. Lee. Robust spotting of key gestures from whole body motion sequence. In *Proc. of the 7th IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG)*, 2006.