Recognition of Deictic Gestures with Context*

Nils Hofemann, Jannik Fritsch, and Gerhard Sagerer

Applied Computer Science
Faculty of Technology, Bielefeld University
33615 Bielefeld, Germany
{nhofeman, jannik, sagerer}@techfak.uni-bielefeld.de

Abstract. Pointing at objects is a natural form of interaction between humans that is of particular importance in human-machine interfaces. Our goal is the recognition of such deictic gestures on our mobile robot in order to enable a natural way of interaction. The approach proposed analyzes image data from the robot's camera to detect the gesturing hand. We perform deictic gesture recognition through extending a trajectory recognition algorithm based on particle filtering with symbolic information from the objects in the vicinity of the acting hand. This vicinity is specified by a *context area*. By propagating the samples depending on a successful matching between expected and observed objects the samples that lack a corresponding context object are propagated less often. The results obtained demonstrate the robustness of the proposed system integrating trajectory data with symbolic information for deictic gesture recognition.

1 Introduction

In various human-machine interfaces more human-like forms of interaction are developed. Especially for robots inhabiting human environments, a multi-modal and human friendly interaction is necessary for the acceptance of such robots. Apart from the intensively researched areas of speech processing that are necessary for dialog interaction, the video-based recognition of hand gestures is a very important and challenging topic for enabling multi-modal human-machine interfaces that incorporate gestural expressions of the human.

In every-day communication deictic gestures play an important role as it is intuitive and common for humans to reference objects by pointing at them. In contrast to other types of gestural communication, for example sign language [10], deictic gestures are not performed independently of the environment but stand in a context to the referenced object. We concentrate on pointing gestures for identifying medium sized objects in an office environment. Recognizing deictic gestures, therefore, means not only to classify the hand motion as *pointing* but also to determine the referenced object. Here we do not consider referencing object details. We will focus on the incorporation of the gesture

* The work described in this paper was partially conducted within the EU Integrated Project COGNIRON ("The Cognitive Companion") funded by the European Commission Division FP6-IST Future and Emerging Technologies under Contract FP6-002020 and supported by the German Research Foundation within the Graduate Program 'Task Oriented Communication'.

C.E. Rasmussen et al. (Eds.): DAGM 2004, LNCS 3175, pp. 334–341, 2004.
© Springer-Verlag Berlin Heidelberg 2004

context, i.e., the referenced object, into a motion-based gesture recognition algorithm resulting in a more robust gesture recognition.

According to Bobick [3], human motion can be categorized into three classes: *movement*, *activity*, and *action*. Each category represents a different level of recognition complexity: A *movement* has little variation in its different instances and is generally only subject to linear scalings, e.g., it is performed at different speeds. An *activity* is described by a sequence of movements but can contain more complex temporal variations. Both, *movement* and *activity* do not refer to elements external to the human performing the motion. Interesting for our view on deictic gestures is the class *action* that is defined by an activity and an associated symbolic information (e.g., a referenced object). Obviously, a deictic gesture 'pointing at object X' can be described with this motion schema. Here, the low level movements are accelerating and decelerating of the pointing hand and the activity is a complete *approach* motion. Combining this activity of the pointing hand with the symbolic data denoting the referenced object X results in recognizing the action 'pointing at object X'. Due to the characteristics of pointing gestures we employ a 2D representation for the hand trajectory based on the velocity and the change of direction of the acting hand in the image.

An important topic for deictic gesture recognition is binding the motion to a symbolic object: During a pointing gesture the hand approaches an object. Using the direction information from the moving hand, an object can be searched in an appropriate search region. If an object is found, a binding of the object to the hand motion can be established. We will show how this binding can be performed **during** processing of the trajectory data resulting in an integrated approach combining sensory trajectory data and the symbolic object data for recognizing deictic gestures with context. We intend to use this recognition system for the multi-modal human-machine interface on-board a mobile robot allowing humans to reference objects by speech and pointing [8].

In this paper we will first discuss related work on gesture recognition in Section 2. Subsequently, we give in Section 3 an overview of the presented system and the used modules. The Particle Filtering algorithm applied for activity recognition is described in Section 4. In Section 5 we show how this algorithm is combined with symbolic object data for recognition of deictic gestures. In Section 6 results of the system acquired in a demonstration scenario are presented, we conclude the paper with a short summary in Section 7.

2 Related Work

Although there is a large amount of literature dealing with gesture recognition, only very few approaches have actually attacked the problem of incorporating symbolic context into the recognition task. One of the first approaches exploiting hand motions and objects in parallel is the work of Kuniyoshi [7] on qualitative recognition of assembly actions in a blocks world domain. This approach features an action model capturing the hand motion as well as an environment model representing the object context. The two models are related to each other by a hierarchical parallel automata that performs the action recognition.

An approach dealing with the recognition of actions in an office environment is the work by Ayers and Shah [1]. Here a person is tracked based on detecting the face and/or neck with a simple skin color model. The way in which a person interacts with an object is defined in terms of intensity changes within the object's image area. By relating the tracked person to detected intensity changes in its vicinity and using a finite state model defining possible action sequences, the action recognition is performed. Similar to Kuniyoshi's approach, no explicit motion models are used.

An approach that actually combines both types of information, sensory trajectory data and symbolic object data, in a structured framework is the work by Moore et al. [9]. Different image processing steps are carried out to obtain *image-based*, *object-based*, and *action-based* evidences for objects and actions. Moore et al. analyze the trajectory of a tracked hand with Hidden-Markov-Models trained offline on different activities related to the known objects to obtain the *action-based* evidence.

Only the approach by Moore et al. incorporates the hand motion, while the approaches by Kuniyoshi and Ayers and Shah rely only on the hand position. However, in the approach of Moore et al. the sensory trajectory information is used primarily as an additional cue for object recognition. We present in the following an approach for reaching the oppositional goal of recognizing gestures with the help of symbolic information.

3 System Overview

Due to the requirements of a fluent conversation between a human and a machine, the system for recognizing deictic gestures has to work in real-time. The overall deictic gesture recognition system is depicted in Fig. 1. The first two modules depicted at the left are designed for operating directly on the image data. The module on the top extracts the trajectory of the acting hand from the video data by detecting skin-colored regions and tracking these region over time (for details see [4], chapter 4). The resulting regions are tracked over time using a Kalman filter with a constant acceleration model. The module at the bottom performs object recognition in order to extract symbolic information about the objects situated in the scene. This module is based on an algorithm proposed by Viola and Jones [11]. In this paper we focus on the action recognition module which contains an activity recognition algorithm that is extended to incorporate symbolic data from the object recognition. In this way, a recognition of deictic gestures with incorporation of their context is realized. The recognition results of the system can facilitate a multi-modal human-machine-interface.

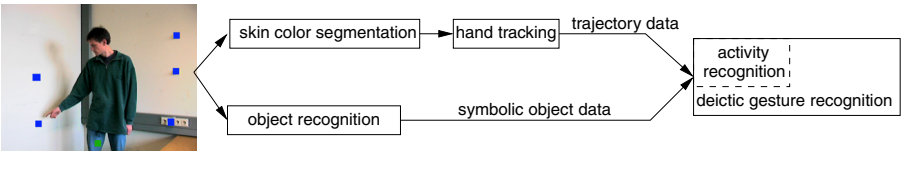

Fig. 1. Architecture of the deictic gesture recognition system.

4 Activity Recognition

Based on the trajectory generated by the acting hand of the human we can classify this trajectory. Since the start and end points of gestures are not explicitly given it is advantageous if the classification algorithm implicitly selects the relevant parts of a trajectory for classification. Additionally, as the same gestures are usually not identically executed the classification algorithm should be able to deal with a certain variability of the trajectory. The algorithm selected for segmentation and recognition of activities is based on the *Conditional Density Propagation* (CONDENSATION) algorithm which is a particle filtering algorithm introduced by Isard and Blake to track objects in noisy image sequences [5]. In [6] they extended the procedure to automatically switch between several activity models to allow a classification of the activities. Black and Jepson adapted the CONDENSATION algorithm in order to classify the trajectories of commands drawn at a blackboard [2].

Our approach is based on the work of Black and Jepson. Activities are represented by parameterized models which are matched with the input data. In contrast to the approach presented by Black and Jepson where motions are represented in an image coordinate system $(\Delta x, \Delta y)$, we have chosen a trajectory representation that consists of the velocity Δr and the change of direction $\Delta\gamma$. In this way we abstract from the absolute direction of the gesture and can represent a wide range of deictic gestures with one generic model. As the user typically orients himself towards the dialog partner the used representation can be considered view-independent in our scenario.

Each gesture model $\mathbf{m}$ consists of a 2-dimensional trajectory, which describes the motion of the hand during execution of the activity.

$$\mathbf{m}^{(\mu)} = \{\mathbf{x}_0, \mathbf{x}_1, \ldots, \mathbf{x}_T\}, \quad \mathbf{x}_t = (\Delta r_t, \Delta\gamma_t) \tag{1}$$

For comparison of a model $\mathbf{m}^{(\mu)}$ with the observed data $\mathbf{z}_t = (\Delta r_t, \Delta\gamma_t)$ the parameter vector $\mathbf{s}_t$ is used. This vector defines the sample of the activity model μ where the time index ϕ indicates the current position within the model trajectory at time t. The parameter α is used for amplitude scaling while ρ defines the scaling in time dimension.

$$\mathbf{s}_t = (\mu_t, \phi_t, \alpha_t, \rho_t) \tag{2}$$

The goal of the CONDENSATION algorithm is to determine the parameter vector $\mathbf{s}_t$ so that the fit of the model trajectory with the observed data $\mathbf{z}_t$ is maximized. This is achieved by temporal propagation of N weighted samples

$$\left\{(\mathbf{s}_t^{(1)}, \pi_t^{(1)}), \ldots, (\mathbf{s}_t^{(N)}, \pi_t^{(N)})\right\} \tag{3}$$

which represent the a posteriori probability $p(\mathbf{s}_t|\mathbf{z}_t)$ at time t. The weight $\pi_t^{(n)}$ of the sample $\mathbf{s}_t^{(n)}$ is the normalized probability $p(\mathbf{z}_t|\mathbf{s}_t^n)$. This is calculated by comparing each scaled component of the model trajectory in the last w time steps with the observed data. For calculating the difference between model and observed data a Gaussian density is assumed for each point of the model trajectory.

The propagation of the weighted samples over time consists of three steps and is based on the results of the previous time step:

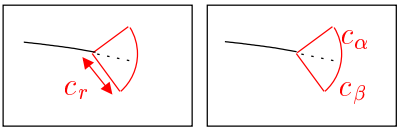

Fig. 2. The definition of the context area.

Select: Selection of N samples $\mathbf{s}_{t-1}^{(n)}$ according to their respective weight $\pi_{t-1}^{(n)}$ from the sample pool at time $t-1$. This selection scheme implies a preference for samples with high probability, i.e., they are selected more often.

Predict: The parameters of each sample $\mathbf{s}_t^{(n)}$ are predicted by adding Gaussian noise to α_{t-i} and ρ_{t-1} as well as to the position ϕ_{t-1} that is increased in each time step by ρ_t. If ϕ_t is larger than the model length $\phi_{\max}$ a new sample $\mathbf{s}_t^{(n)}$ is initialized.

Update: Determination of the weights $\pi_t^{(n)}$ based on $p(\mathbf{z}_t|\mathbf{s}_t^{(n)})$.

Using the weighted samples obtained by these steps the classification of activities can be achieved. The probability that a certain model μ_i is completed at time t is given by its so-called end-probability $p_{\text{end}}(\mu_i)$. This end probability is the sum of all weights of a specific activity model with $\phi_t > 0.9\phi_{\max}$.

For the overall recognition system the repertoire of activities consists of *approach* and *rest*. The model *rest* is used to model the time periods where the hand is not moving at all. With these models the trajectory-based recognition of deictic gestures can be performed.

5 Recognition of Pointing Actions

As mentioned in the introduction a deictic gesture is always performed to reference an object more or less in the vicinity of the hand. To extract this fundamental information from the gesture, both the movement of the hand represented by the trajectory and symbolic data describing the object have to be combined. This combination is necessary if several objects are present in the scene as only using the distance between the hand and an object is not sufficient for detecting a pointing gesture. The hand may be in the vicinity of several objects but the object referenced by the pointing gesture depends on the direction of the hand motion. This area where an object can be expected in the spatial context of an action is called *context area*.

In order to have a variable context area we extend the model vector $\mathbf{x}_t$ (Eq. 1) by adding parameters for this area. It is defined as a circle segment with a search radius c_r and a direction range, limited by a start and end angle (c_α, c_β). These parameters are visualized in Fig. 2. The angles are interpreted relative to the direction of the tracked hand. The *approach* model consists of some time steps with increasing velocity but without a context area in the beginning later in the model a context area is defined with a shrinking distance c_r and the hand slows down.

To search objects in a context area relative to the hand position the absolute position (P_x, P_y) of the hand is required. According to this demand the complete input data consists of the observed motion data $\mathbf{z}_t$ and the coordinates P_x, P_y.

The spatial context defined in the models is incorporated in the CONDENSATION algorithm as follows. In each time-step the trajectory and context data is sequentially processed for every sample. At first the values of the sample are predicted based on the activity of the hand, afterwards the symbolic object data in relation to the hand is considered:

If there are objects in the context area of the sample at the current time index ϕ_t one object is selected randomly. For adding this symbolic data to the samples of the CONDENSATION we extend the sample vector s_t (Eq. 2) by a parameter ID_t denoting a binding with a specific object:

$$s_t = (\mu_t, \phi_t, \alpha_t, \rho_t, \mathrm{ID}_t) \tag{4}$$

This binding is performed in the *Update* step of the CONDENSATION algorithm. An object found in the context area is bound to the sample if no binding has occurred previously. Once the the sample s_t contains an object ID it will be propagated with the sample using $\mathrm{ID}_t^{(n)} = \mathrm{ID}_{t-1}^{(n)}$.

Additional we extend the calculation of the sample weight with a multiplicative *context factor* P_{symb} representing how good the bound object fits the expected spatial context of the model.

$$\pi_t^{*(i)} \propto p(\mathbf{z}_t|\mathbf{s}_t^{(i)})\ P_{symb}(\mathrm{ID}_t|\mathbf{s}_t^{(i)}) \tag{5}$$

For evaluating pointing gestures we use a constant factor for P_{symb}. The value of this factor depends on whether a previously bound object (i.e., with the correct ID) is present in the context area or not. We use P_{symb} = 1.0 if the expected object is present and a smaller value $P_{symb} = P_{missing}$ if the context area does not contain the previously bound object. This leads to smaller weights $\pi_t^{*(i)}$ of samples with a missing context so that these samples are selected and propagated less often.

When the threshold for the end probability $p_{end}^{(i)}$ for one model is reached the parameter ID is used for evaluating the object the human pointed at. One approach is to count the number of samples bound with an object. But this is an inaccurate indicator as all samples influence the result with the same weight. Assuming a large number of samples is bound with one object but the weight of these samples is small this will lead to a misinterpretation of the bound object. A better method is to select an object bound to samples with a high weight, as the weight of a sample describes how good it matches the trajectory in the last steps. Consequently, we calculate for each object O_j the sum p_{O_j} of the weights of all samples belonging to the recognized model μ_i that were bound to this object.

$$p_{O_j}(\mu_i) = \sum_{n=1}^{N} \begin{cases} \pi_t^{*(n)} & \text{, if } \mu_i \in \mathbf{s}_t^{(n)} \wedge (\phi_t > 0.9\phi_{\max}) \wedge \mathrm{ID}_t = O_j \\ 0 & \text{, else} \end{cases} \tag{6}$$

If the highest value $p_{O_j}(\mu_i)$ for the model is larger than a defined percentage ($T_O = 30\%$) of the model end probability $p_{\mathrm{end}}(\mu_i)$ the object O_j is selected as being the object that was pointed at by the 'pointing' gesture. If the model has an optional spatial context and for all objects the end probability $p_{O_j}(\mu_i)$ is lower than required the model is recognized without an object binding.

The benefit of the described approach is a robust recognition of deictic gestures combined with information about the referenced object. The system is able to detect not only deictic gestures performed in different directions but also provides the object the human pointed at.

6 Results

We evaluated the presented system in an experimental setup using 14 sequences of deictic gestures executed by five test subjects resulting in 84 pointing gestures. An observed person stands in front of a camera at a distance of approximately 2m so that the upper part of the body and the acting hand are in the field of view of the camera. The person points with the right hand at six objects (see Fig. 1), two on his right, three on his left side, and one object in front of the person. We assumed perfect object recognition results for the evaluation. For this evaluation only the localization of objects was needed, as *pointing* is independent of a specific object type. The images of size 320x240 pixels are recorded with a frame-rate of 15 images per second. In our experiments real-time recognition was achieved using a standard PC (Intel, 2.4GHz) running with Linux. The models were built by averaging over several example gestures.

In the evaluation (see Tab. 1) we compare the results for different parameterizations of the gesture recognition algorithm. For evaluation we use not only the recognition rate but also the word error rate (WER) which is defined by WER [1] $= \frac{\#I+\#D+\#S}{\#E}$. As parameters for the CONDENSATION we use N=1000 samples, the scaling factors α and ρ are between 0.65 and 1.35 with variance $\sigma = 0.15$.

Table 1. Recognition of deictic gestures

	Context								
	none	distance	directed	weighted					
$P_{missing}$	-	1.0	1.0	0.8	0.6	0.4	0.2	0.1	0.0
Correct	83	69	74	72	75	77	76	78	82
Insertion	81	9	5	5	5	5	6	5	18
Deletion	1	10	10	12	9	7	6	6	2
Substitution	0	5	0	0	0	0	0	0	0
Word error rate	97.6	28.6	17.8	20.2	16.7	14.3	14.3	13,3	23.8
Recognition rate	98.8	82.2	88.1	85.7	89.3	91.7	90.4	92.8	97.6

The second column (*'none'*) shows the results with the standard trajectory-based approach of Black et al. [2]. Without incorporation of the symbolic context no separation between departing and approaching activities is possible, every straight motion is interpreted as *pointing*. Therefore, this approach gives the highest recognition rate but it also results in the highest WER due to a huge number of insertions. Note that there is also no information about which object is referenced by the pointing gesture.

[1] using I:Insertion, D:Deletion, S:Substitution, E:Expected

By using the distance (column *'distance'*) between the approaching hand and the surrounding objects mainly gestures approaching an object are recognized. But still a high rate of insertions and even substitutions (i.e., a wrong object binding) is observed. The substitutions show the disadvantage of a simple distance criterion that does not incorporate the direction of the hand motion.

Using a directed context area (column *'directed'*) we achieve a better recognition rate and a lower WER. By introducing a weighting (columns *'weighted'*) for samples not matching the expected context, the recognition rates can be further increased while reducing the WER. If samples not matching the context are deleted ($P_{missing} = 0$) the recognition rate is further increased but now also the WER is increased. This is due to the fact that all samples with a missing context area are deleted and indirectly those samples not matching the trajectory but with a bound object are propagated.

7 Summary

In this paper we presented an integrated approach to deictic gesture recognition that combines sensory trajectory data with the symbolic information of objects in the vicinity of the gesturing hand. Through the combined analysis of both types of data our approach reaches an increased robustness within real time. The recognition result provides not only the information that a deictic gesture has been performed, but also the object that has been pointed at.

References

1. D. Ayers and M. Shah. Monitoring human behavior in an office environment. In *IEEE Workshop on Interpretation of Visual Motion, CVPR-98*, Santa Barbara, CA, June 1998.
2. M. J. Black and A. D. Jepson. A probabilistic framework for matching temporal trajectories: CONDENSATION-based recognition of gestures and expressions. *Lecture Notes in Computer Science*, 1406:909–924, 1998.
3. A. Bobick and Y. Ivanov. Action recognition using probabilistic parsing. In *Proc. of CVPR*, pages 196–202, Santa Barbara, California, 1998.
4. Jannik Fritsch. *Vision-based Recognition of Gestures with Context*. Dissertation, Universität Bielefeld, Technische Fakultät, 2003.
5. M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. *Lecture Notes in Computer Science*, 1064:343–356, 1996.
6. M. Isard and A. Blake. A mixed-state condensation tracker with automatic model-switching. In *ICCV'98*, pages 107–112, Mumbai, India, 1998.
7. Y. Kuniyoshi and H. Inoue. Qualitative recognition of ongoing human action sequences. In *Proc. International Joint Conference on Artificial Intelligence*, pages 1600–1609, 1993.
8. Frank Lömker and Gerhard Sagerer. A multimodal system for object learning. In Luc Van Gool, editor, *Pattern Recognition, 24th DAGM Symposium, Zurich, Switzerland*, Lecture Notes in Computer Science 2449, pages 490–497, Berlin, September 2002. Springer.
9. D. Moore, I. Essa, and M. Hayes. Exploiting human actions and object context for recognition tasks. In *Proceedings of IEEE Int. Conf. on Computer Vision*, Corfu, Greece, 1999.
10. T. Starner and A. Pentland. Visual recognition of american sign language using hidden markov models. In *Int. Workshop on Automatic Face and Gesture Recognition*, 1995.
11. P. Viola and M. Jones. Robust real-time object detection. In *Proc. IEEE Int. Workshop on Statistical and Computational Theories of Vision*, Vancouver, Canada, 2001.