FP6-IST-002020

COGNIRON

The Cognitive Robot Companion

Integrated Project

Information Society Technologies Priority

D2.2.1
Report on definitions of human attribute model and posture estimation

Due date of deliverable: 31.12.2004
Actual submission date: 31.01.2005

Start date of project: January 1st, 2004 **Duration** : 48 months

Organisation name of lead contractor for this deliverable: University of Karlsruhe

Revision: Final
Dissemination level: PU

Executive Summary

This report summarises our research results regarding human attribute model and posture estimation for a mobile robot companion. First of all a short view over the geometric model and the data structure all partners of RA2 decided to use will be given. Next a summary of human kinematics will be given, which, together with the geometric model, form the human attribute model. The report goes on with a summary on our results on pose estimation, which is based on some published or internal scientific reports. First the results on deictic gesture recognition based on 2D information will be presented. A skin colour based hand tracking was developed to estimate trajectories to objects. This algorithm can operate with a single 2D colour camera. Afterwards the results on 3D pose estimation of the upper body part will be summarised. In this approach the geometric human model is used to estimate the positions of body parts in data from 3D sensors. The scientific reports referred to are attached to this report.

Role of human attribute model and posture estimation in Cogniron

Communication with humans is essential for a cognitive robot companion and one way of natural communication for humans are gestures. Therefore a robot companion must be able to recognise and understand gestures. In addition to communication interaction and learning are two essentials. For both a robot needs to be aware of what are the humans doing how. The knowledge of human attributes and kinematics as well as algorithms to estimate the positions of human body parts are bases for these functionalities. This report describes a solid base for these cognitive abilities.

Relation to the Key Experiments

The here presented results are important for all key-experiments. In KE1 “Robot Home Tour” the human uses deictic gestures to show and tell the robot objects in the different rooms. In KE2 “Curious Robot” the robot should observe the humans activities and therefore needs to estimate the humans posture over time. This is also important for KE3 “Skill and Task Learning” where the robot should learn new abilities from observing the human’s motions and activities. The human attribute model helps the robot to map the observed motions to its own kinematics.

1 Human attribute model and posture estimation

Starting with the human attribute model the geometric human model and its data structure will be shown. In addition to this, the human kinematics and their use for estimating information will be explained. The second part will then give a summary of the published scientific paper and the internal reports regarding the posture estimation.

1.1 Geometric human model

The geometric human model the partners of RA2 decided on is a cylinder-based model, which can be adapted to the proportions of the represented human. Therefore all cylinders are "degenerated" (also see [2]), which means it has two parallel ellipses having the same rotation. These cylinders are described with five parameters, which are:

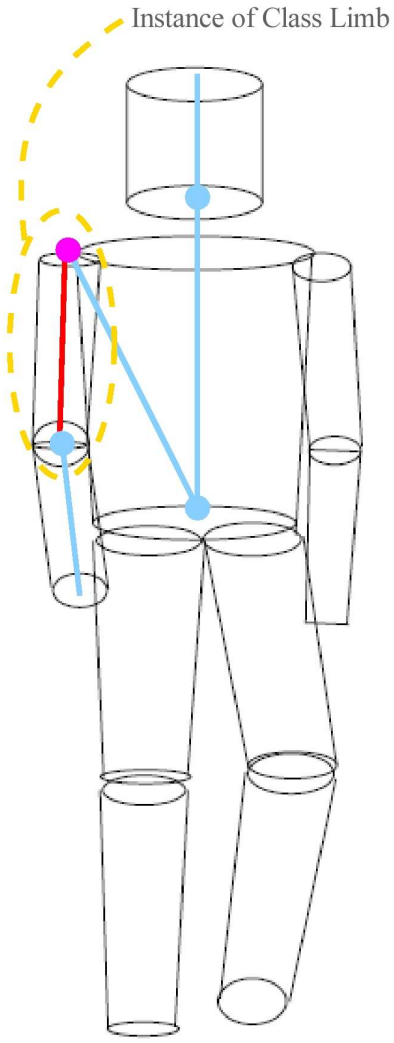
- Major-semi axis of ellipsis at cylinder beginning
- Minor-semi axis of ellipsis at cylinder beginning
- Major-semi axis of ellipsis at cylinder end
- Minor-semi axis of ellipsis at cylinder end
- Length of cylinder

The basic model consists of, but is not limited to ten cylinders:

- Two for upper arms
- Two for lower arms
- Two for upper legs
- Two for lower legs
- One for the torso
- One for the head

The geometric human model is shown in picture 1.

Picture 1: Geometric human model

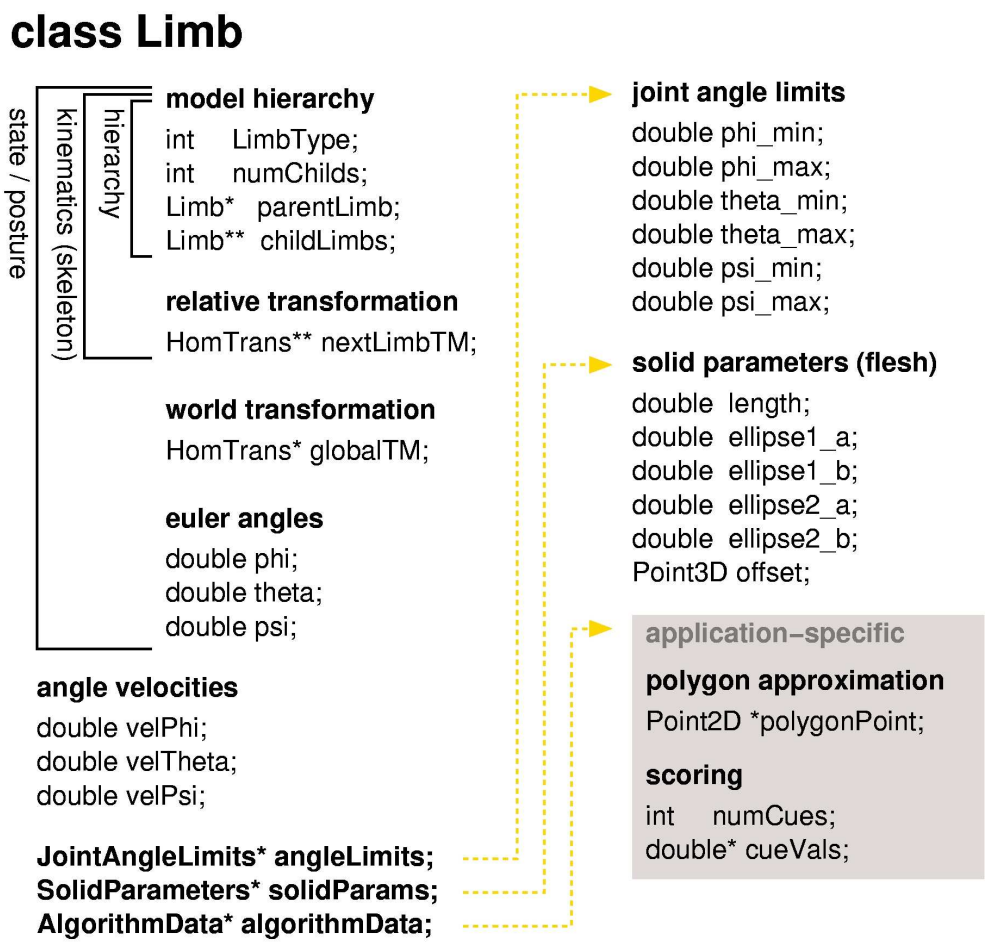


1.2 Data structure for geometric human model

For all future work within the consortium it was necessary to decide on a uniform data structure for the geometric human model described in 1.1. This data structure should enable the exchange and combination of algorithms within the consortium without the need to adapt interfaces. The structure is build up in a way, that all limbs of the model are linked to a parent limb. The only exception is the base limb, namely the torso, which all other limbs are directly or indirectly connected to. This structure makes it possible to add limbs if necessary, e.g. one limb for the hand, one for each finger or three for each finger.

Each limb contains information about the parent limb, potential child limbs, its own position and orientation relative to the parent limb or in world coordinates. To put flesh on the bones each limb has the five parameters for the cylinders described in 1.1. There is also space reserved for information for specific algorithms like maximum possible angles. An overview of the data structure is shown in picture 2.

Picture 2: Data structure for human model



1.3 Human kinematics

In order to get a robust estimation of the pose of a human in an all day environment it is necessary to have some basic knowledge of human kinematics. In [3] the important joints with their maximum angles, as well as all aspect ratios of human limbs are described. The geometric human model described in 1.1 covers all these joints and limbs. To get good results for the maximum number of humans, only average values have been taken. Higher maximum angles from taking exercises or lower angles from injury or diseases are not considered. In a later phase the angles will be adaptive, so that every human gets an own set. Not every junction has three degrees of freedom as the data structure of the model provides. Regarding the degrees of freedom of a joint, one or two angles are set to zero. Other important aspects of human kinematics are the average measurements of body parts and the aspect ratio between body height and body parts. These measurements can also be found in [3]. It is interesting and very useful, that the aspect ratios stay the same at different body heights. Slight differences can be found between the aspect ratios of male and female humans. This might become useful for distinguishing between the genders.
A glossary for the Greek and Latin terms can be found in [4].

1.4 Detection of deictic gestures

As pointing gestures are important for key-experiment 1 “Robot Home Tour” the detection of these gestures was separated from the 3D pose estimation at first. The in [1] presented approach analyses

data from a 2D colour camera to detect the hands of a human. The deictic gesture is recognised through the extension of a trajectory recognition algorithm based on particle filtering with symbolic information from the objects in the vicinity of the pointing hand. The size of the vicinity is adapted by the context to get optimal results. Latest optimisations brought the processing speed of the algorithm presented in [1] on a normal PC up to 15 frames per second. Another project of one partner deals with an approach of detecting deictic gestures in 3D from monocular colour images described in [5]. The scientific outcome of this project is also useful for the 3D pose estimation in Cogniron.

1.5 3D pose estimation

For the estimation of human poses in 3D the iterative closest point algorithm is used to fit the geometric human model described in 1.1 to 3D sensor data. This algorithm finds a rotation and translation for the cylinders such that the error between the model and the 3D data points becomes minimal. Therefore corresponding points for the data points must be found on the surface of the cylinders. Checking all points with all cylinders finds the corresponding cylinder to a set of data points. This check and the iterative closest point algorithm are performed iteratively. First results of this algorithm with the upper part of the body model (consisting of 6 cylinders) can be found in [2].

2 Future work

The results of the approaches presented in [1] and [2] showed, that it is possible to estimate the pose of a human with one sensor in really simple environments. As in Cogniron a robot companion needs to operate very robust in all day environments, these approaches have to be combined into a multi modal gesture recognition based on both 2D and 3D data. Also the algorithms itself need to be tested and optimised more, as 7.5 frames per second for detection is quiet good for now, but still not the end of the road. In [1] only one arm, in [2] only the upper body parts are detected. In the future the whole body has to be taken into account for pose estimation, because gesture recognition and activity detection need to see the whole human in order to give a correct interpretation.
Future work on the human model will focus on the kinematics and the verification of the outputs of the pose estimation algorithms. With the maximum angles and angular speeds of body parts the estimated positions can be verified and corrected if necessary. These angles and speeds, first fixed for all individuals, will be adapted over time for humans in the robots daily environment. The knowledge of the human anatomy and the relations between body parts will be used to estimate size and position of parts that are out of the robots sight.

3 References

[1] N. Hofemann and J. Fritsch and G. Sagerer. Recognition of Deictic Gestures with Context. In DAGM04, pages 334-341, Springer-Verlag, Heidelberg, Germany, 2004

[2] S. Vacek. 3D pose estimation of human body. *University of Karlsruhe internal report*, 2004

[3] A. Freisinger and K. Pfeiffer. Values of human kinematics. *IPA internal report*, 2004

[4] A. Freisinger and K. Pfeiffer. Glossary of human kinematics. *IPA internal report*, 2004

[5] J. Schmidt and J. Fritsch and G. Sagerer. Real-time 3d hand and arm tracking from monocular color images. Draft for ICIP, 2005

Appendix

Attached scientific paper and internal reports (see references)

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'.

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.

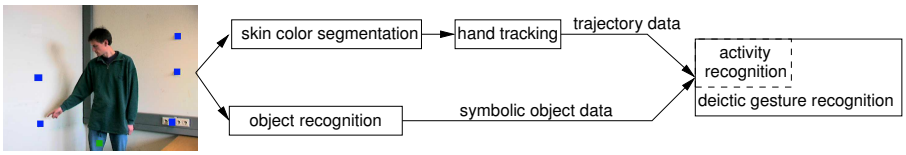


Fig. 1. Architecture of the deictic gesture recognition system.

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:

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

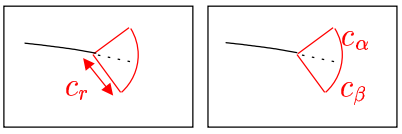


Fig. 2. The definition of the context area.

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 $^\star = \frac{\#I+\#D+\#S}{\#E}$. 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 $\sigma = 0.15$.

	Context								
	none	distance	directed	weighted					
$P_{missing}$	-	1.0	1.0	0.8	0.6	0.4	0.2	0.1	0.0
Correct	82	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

Table 1. Recognition of deictic gestures

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.

* 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.

Description of 3D Pose Estimation at UniKarl

In order to estimate the 3d pose of a human being, UniKarl uses the ICP (Iterative Closest Point) algorithm to fit an artificial 3d human model to the 3d sensor data of the stereo camera.

At the moment the 3d human model consists of 6 body parts (torso, head, left upper and lower arm, right upper and lower arm) which are connected through 5 joints (neck, left and right shoulder, left and right elbow). According to the 3d human model developed in RA2, each body part is described by a "degenerated" cylinder, which is a field described by two parallel ellipses having the same rotation. The following figure 1 tries to give an idea of how such a field looks like. Figure 2 shows the overall 3d human body model.

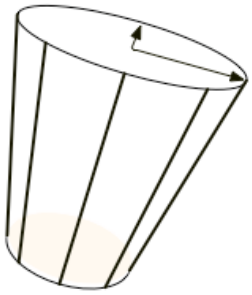


Figure 1: "degenerated" cylinder

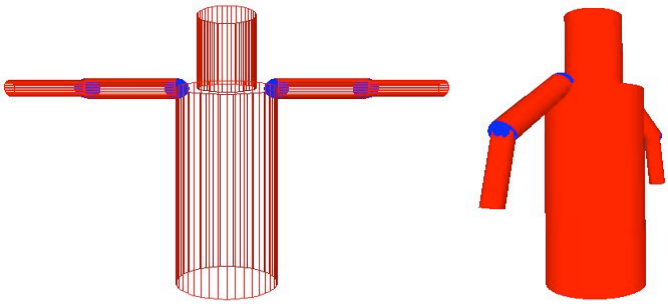


Figure 2: resulting 3d human body model

In general, the ICP-approach is based on the best fit of a 3d geometrical model of the human body to a set of 3d data points. These point clouds are acquired using the stereo camera device (see report of WP 2.1). Figure 3 shows the basic idea of the ICP.

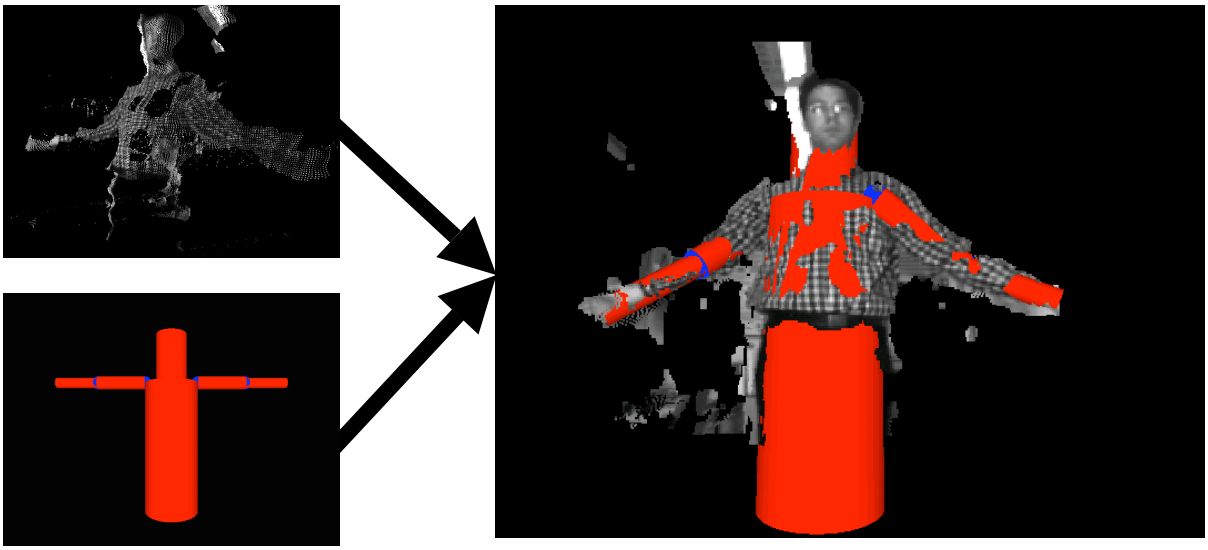


Figure 3: Fitting 3d camera data to the 3d human body model

The basic idea of the iterative-closest-point algorithm is to find a rotation and translation such that the overall error between the model and the data points becomes minimal:

$$f(R,\vec{t}) = \frac{1}{N}\sum_{i=1}^{N} || R(\vec{x_i}) + \vec{t} - \vec{p_i} ||^2$$

Here, $X = \{\vec{x}_i\}$ describes the points of the model and $P = \{\vec{p}_i\}$ describes the data points of the sensor. For each i, $\vec{p}_i$ corresponds to $\vec{x}_i$ and the ICP-algorithm searches for a rotation R and translation $\vec{t}$ in order to minimize the above equation.

The most crucial (and computationally most expensive) part is to set up the two lists of corresponding points. The problem is, that the model is not described by a set of points but rather by a geometrical field. To establish the correspondence between a data point and the model, the *closest* point of the model is chosen as depicted in figure 4. There are some special cases which have to considered separately, e.g. the data point lies over the "degenerated" cylinder. Data points, which are too far away from the field, are not taken into account because they should not have any influence on the resulting rotation and translation.

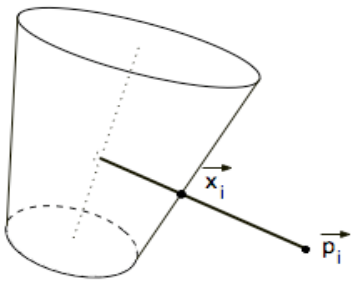


Figure 4: Finding the closest point on the "degenerated" cylinder

The 3d human body model is composed out of several body parts and for each body part, the ICP is used separately. To construct the point lists of closest points, each point is tested with each body part and the nearest body part is chosen as the corresponding part. If a point has equal distance to several body parts, the point is considered for each of these. After applying the icp, an additional test has to be made in order to ensure, that the body parts are still connected. This test is done by checking the position of the start point of the one body part with the end position of the other body part.

The algorithm is an iterative one. In each cycle, the position of the model is improved until the change of the position is small. For each cycle, the configuration of the 3d human body model of the previous cycle is used. When the algorithm is called for the first time, the model configuration of the previous frame is used. Summarizing, the algorithm works as follow:

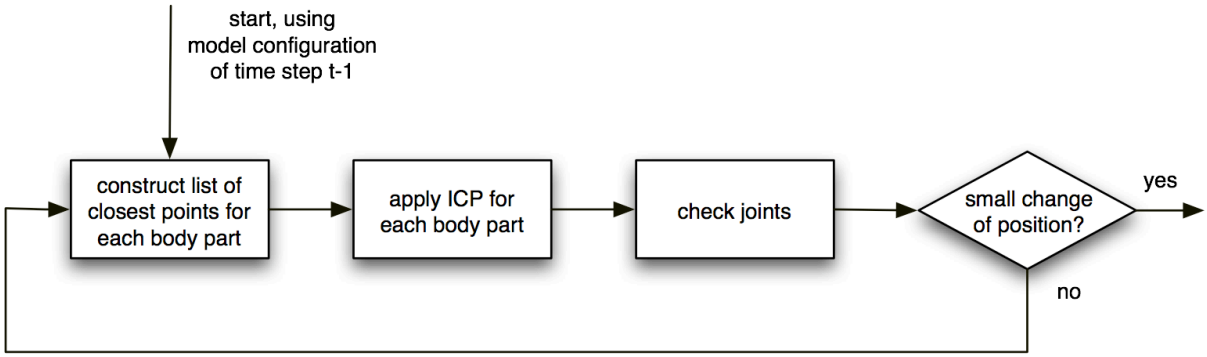


Figure 5: Overview of on cycle of the algorithm

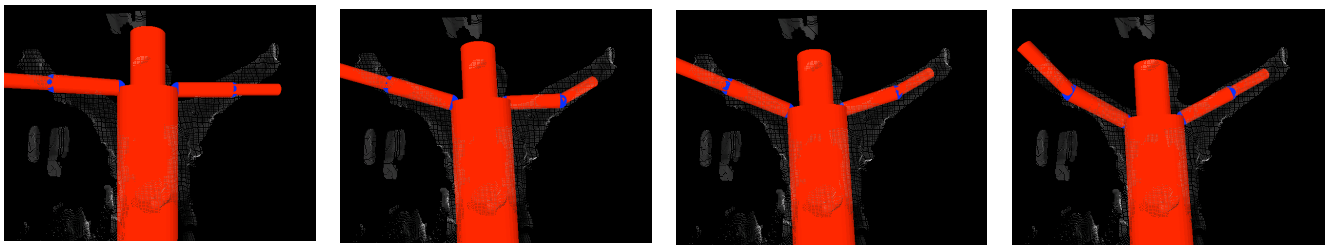


Figure 6: Iterative fitting of the model to the sensor data

Some first results can be seen in the two figures 6 and 7. The first row (figure 6) shows the iterative approaching of the model to the 3d sensor data. As can be seen in the first frame, the model isn't well conditioned but with a few ongoing iterations the model adapts very well to the sensory data.

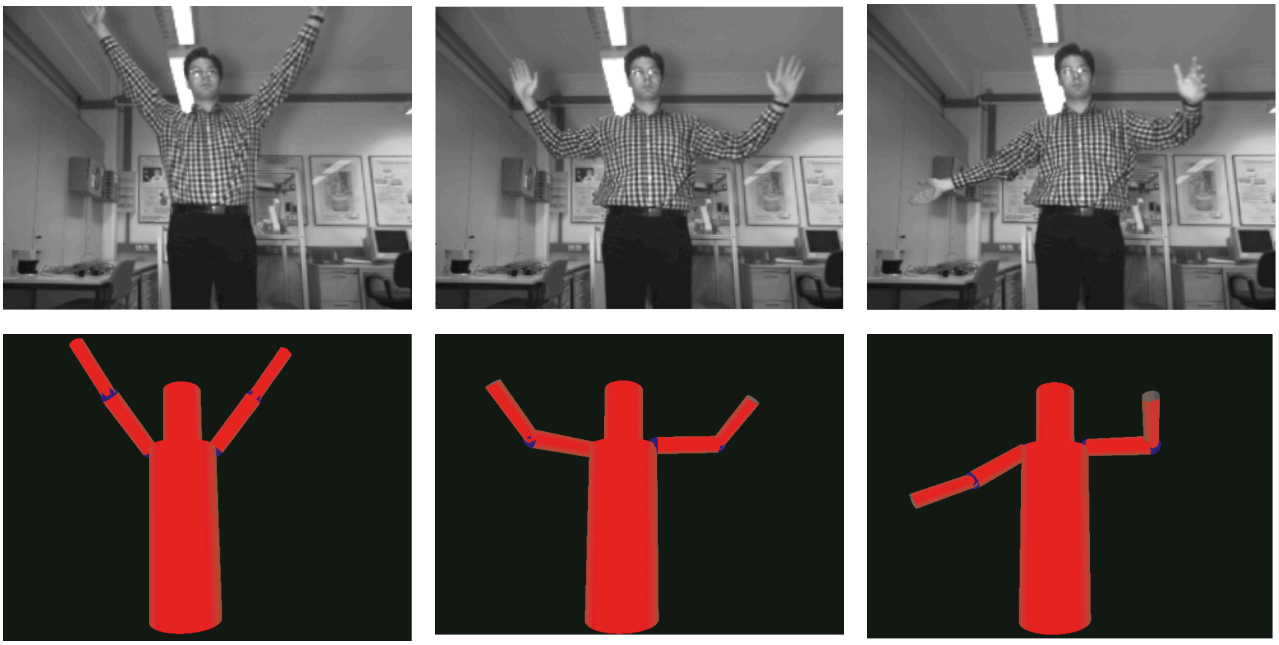


Figure 7: Frames 70, 100, 130 out of a sequence.

The next pictures (figure 7) show three frames out of a sequence together with the corresponding estimated pose. Again, the pose estimation works well. Each frame took 200ms-500ms on a Pentium IV, 30ms were needed for calculating the depth image. Most of the time was used for establishing the closest point relation of each data point with its corresponding cylinder. Beside the needed computing power, another problem is the tracking of the arms. If an arm points towards the camera or if an arm is along the torso, the arm is lost and has to be re-initialised. In order to make the tracking of the arms more robust, additional features will be taken into account like skin-colour for tracking the hands separately and feeding this additional information into the pose estimation.

Values of human kinematics

Andrea Freisinger and Kai Pfeiffer

Fraunhofer IPA
70569 Stuttgart, Germany

1 Introduction

This report presents the outcome of a research of human kinematics. It was performed in the course of a European project called Cogniron. It builds a base for a human attribute model and a verification for pose estimating algorithms. As human physiognomy is an own field of science, the first challenge was, to find out what parts of the kinematics are important and to what level of detail. Due to the fact that estimation of postures relies on visual sensing it is clear, that only the visible joints and the visible parts of the body are interesting. In a geometric human model within Cogniron some of the joints were reduced to single ball joints, even if they consist of several different types of joints like the shoulder, which consists of shoulder joint, shoulder girdle and collarbone. After specifying the important joints the maximum angles for all these joints where determined. The complete set of angles can be found in 2.
Other important aspects of human kinematics are the average measurements of body parts and the aspect ratio between body height and body parts. These measurements can be found in 3. It is interesting and very useful, that the aspect ratios stay the same at different body heights. Slight differences can be found between the aspect ratios of male and female humans. This might become useful for distinguishing between the genders.
A glossary for the Greek and Latin terms can be found in [1].

2 List of maximum joint angles

Name of joint	Type of motion	Max. angle	remarks
Cervical spine	Lateral flexion	35°	to all sides
	Vental flexion	65°	
	Dorsal flexion	40°	
	Rotation	50°	to all sides
Thoracic and lumbar spine	Lateral flexion	40°	to all sides
	Vental flexion	85°	
	Dorsal flexion	60°	
	Rotation	40°	to all sides
Spine (as whole)	Lateral flexion	75°	to all sides
	Vental flexion	150°	
	Dorsal flexion	100°	
	Rotation	90°	to all sides
Shoulder joint with clavicle joint	Flexion	170°	
	Extension	40°	
	Abduction	180°	
	Adduction	40°	
	Inside rotation	100°	
	Outside rotation	90°	
Elbow joint	Flexion	150°	

	Extension	10°
	Pronation	90°
	Supination	90°
Wrist	Dorsal extension	40° - 60°
	Palmar refelexion	60° - 80°
	Radial abduction	20°
	Ulnar abduction	30° - 40°
Hip joint (standing)	Flexion	140°
	Extension	20°
	Abduction	50°
	Adduction	30°
	Inside rotation	40°
	Outside rotation	30°
Hip joint (at 90° defection)	Abduction	80°
	Adduction	20°
	Inside rotation	40°
	Outside rotation	50°
Knee joint (stretched)	Flexion	5° - 10°
	Extension	120° - 150°
Knee joint (at 90° defection)	Inside rotation	10°
	Outside rotation	30° - 40°
Ankle joint (standing foot)	Plantar flexion	30°
	Dorsal extension	50°
	Eversion	10°
	Inversion	20°
Ankle joint (hanging foot)	Plantar flexion	40° - 50°
	Dorsal extension	20° - 30°

3 Lists of anatomic measurements

3.1 Explanation of measurements

This sub chapter gives an explanation of how the measurements are taken, which are stated in 3.2 and 3.3. The definitions are taken from DIN 33402 and divided into standing and sitting. The different columns in 3.2 and 3.3 titled with P1, P5, P50, P95 and P99 show the statistic distribution of measurements in mm. The value at P5 for example means that for 5% of the population the measurement is lower and for 95% it is higher. The column % from kph gives the percentage of the measurement compared to the over all body height. This value is nearly independent from the body height and therefore builds a good base for measurement estimations.

Measurements standing:

Name of measurement	**Explanation**	**Abbreviation**
Body height	Vertical distance from the platform to the highest point of the head in the median plane.	Kph
Eye height	Vertical distance from the platform to the inner angle of the right eye.	auh
Chin height	Vertical distance from the platform to the lower point of the chin.	Kih
Acromial shoulder height	Vertical distance from the platform to the lateral most jutting out point of the shoulder height of the right omo-	sha

height	plate.	
Elbow height	Vertical distance from the platform to the lowest point of the perpendicular bend right elbow.	ebh
Knee height	Vertical distance from the platform to the proximal furthermost point of the median side of the right shinbone.	kgh
Foot height	Vertical distance from the platform to the distal most jutting out point of the medial ankle of the right foot.	Fh
Span width of the arms	Maximum horizontal distance between the distal furthermost points of the middle fingers with maximum sideways stretched arms and hands.	spwa
Head height	Projective vertical distance from the highest point of the vertex in the median plane (Symmetry plane of the body) to the lowest edge of the lower jaw in the median plane.	koh
Pupil distance	Straight-line distance between the center of both pupils.	Pd
Arm length (additive)	Additive measurement: Sum of upper arm length (oal), forearm length (ual) und hand length (hdl).	Ala
Upper arm length	Straight-line distance from the lateral most jutting out point of the shoulder height of the right omoplate to the proximal furthermost point of the radius head.	Oal
Forearm length	Straight-line distance between the proximal furthermost point of the right radius head and the distal furthermost point of the gripping appendix of the radius.	Ual
Hand length	Straight-line distance from the center of a dorsal connecting line between the distal points of both wrist ankles to the distal furthermost point of the right middle finger of the stretched right hand.	Hdl
Middle finger length	Straight-line distance from the proximal bend groove of the base joint of the right middle finger to the distal most jutting out point of the tip of the right middle finger.	Mfl
Thumb length	Straight-line distance from the base of the right thumb to the distal most jutting out point of the tip of the stretched right thumb.	Dlf
Projective leg length	Differential measurement: Body height (kph) – Fundament-head length (stl)	Blp
Projective thigh length	Differential measurement: Projective leg length (blp) – knee height (kgh)	oslp
Lower leg length	Differential measurement: knee height (kgh) – foot height (fh)	Usl
Foot length	Straight-line distance from the hindmost point of the right heel to point of the first or second toe, which is the furthermost with standing foot.	fl
Head width	Largest horizontal width of the head in a frontal plane perpendicular to the median plane, meaning the straight-line distance between the two lateral most jutting out points at the side of the head in a frontal plane.	kob
Shoulder width (bideltoidale)	Horizontal distance between the two lateral most jutting out points of the shoulder height of the lateral shoulder contour forming delta structure.	Sbd
Elbow width	Straight-line distance between the two furthermost points of the joint knots of the right upper arm bone.	Ebb
Direct hand width	At the right hand measured straight-line distance from the lateral furthermost point of middle hand bone II in the area of the forefinger base limb to the lateral furthermost point of the middle hand bone V in the area of the little finger base limb of the stretched right hand.	Hdb
hand width with	Perpendicular to the hand axis measured projective	Hdbd

thumb	distance between the lateral furthermost point of the middle hand bone I in the area of the thumb base limb and the lateral furthermost point of the middle hand bone V in the area of the little finger base limb of the stretched right hand.	
Hip width	Largest horizontal distance between the two lateral furthermost points of the hip.	Hueb
Knee width	Straight-line distance between the two joint knots of the right thigh bone with the largest distance while the knee is bend.	Kb

Measurements sitting:

Name of measurment	**Explanation**	**Abbreviation**
Fundament-head length	Vertical distance from the seat to the highest point of the vertex in the median plane.	Stl
Eye height (sitting)	Vertical distance from the seat to the inner eye angle of the right eye.	Auhs
Acromial shoulder height (sitting)	Vertical distance from the seat to the lateral furthermost point of the shoulder height of the right omoplate.	shas
Elbow height (sitting)	Vertical distance from the seat to the lowest point of the perpendicular bend right elbow.	ebhs
Fundament knee length	From the backrest measured horizontal distance of the dorsal furthermost point in the area of the fundament to the distal furthermost point of the right knee at the lower end of the kneecap.	gkl
Knee height (sitting)	Vertical distance from the platform to the highest point on the upper side of the right thigh bend perpendicular to the lower leg in the area of the joint knots of the thigh behind the upper edge of the kneecap. Measured while sitting.	knh
Lower leg length with foot	Vertical distance from the platform to the lower side of the right thigh bend perpendicular to the lower leg right behind the hollow of the knee, measured while sitting Meaning the distance between the platform and the seat.	usfl

3.2 Values female

Name of measurment	**Abbr.**	**P 1**	**P 5**	**P 50**	**P 95**	**P 99**	**% from kph**
Body height	kph	1537	1576	1680	1770	1804	100,00
Eye height	auh	1433	1471	1568	1655	1695	93,35 ± 0,13
Chin height	kih	1330	1368	1471	1553	1595	87,16 ± 0,60
Acromial shoulder height	sha	1249	1279	1374	1454	1502	81,59 ± 0,50
Elbow height	ebh	936	960	1039	1107	1133	61,55 ± 0,82
Knee height	kgh	394	411	453	493	512	26,63 ± 1,11
Foot height	fh	53	59	72	84	90	04,06 ± 0,65
Span width of the arms	spwa	1502	1534	1659	1775	1819	98,52 ± 1,47
Head height	koh	182	188	205	223	228	12,14 ± 0,38
Pupil distance	pd	52	54	60	65	67	03,51 ± 0,14
Arm length (additive)	ala	656	672	731	784	804	43,28 ± 0,83

Upper arm length	oal	274	282	310	336	350	18,29 ± 0,58
Forearm length	ual	214	219	243	265	275	14,31 ± 0,54
Hand length	hdl	155	162	178	194	198	10,48 ± 0,44
Middle finger length	mfl	64	68	75	83	85	04,41 ± 0,26
Thumb length	dlf	53	56	66	76	81	03,81 ± 0,42
Projective thigh length	oslp	275	294	337	380	394	19,52 ± 1,79
Lower leg length	usl	319	339	379	419	432	22,12 ± 1,46
Foot length	fl	221	227	243	263	270	14,53 ± 0,24
Head width	kob	136	139	147	156	161	08,81 ± 0,05
Shoulder width (bideltoidale)	sbh	360	371	402	447	467	24,04 ± 0,92
Elbow width	ebb	53	56	61	67	70	03,60 ± 0,17
Direct hand width	hdb	67	70	76	82	86	04,49 ± 0,14
hand width with thumb	hdbd	80	84	92	101	105	05,43 ± 0,25
Hip width	hueb	303	311	350	392	405	20,61 ± 1,22
Knee width	kb	71	76	86	95	99	04,98 ± 0,37
Fundament-head length	stl	823	840	891	941	960	53,26 ± 0,26
Eye height (sitting)	auhs	711	733	779	829	854	46,49 ± 0,29
Acromial shoulder height (sitting)	shas	510	542	583	634	647	34,52 ± 1,32
Elbow height (sitting)	ebhs	190	203	242	286	302	13,95 ± 1,90
Fundament knee length	gkl	530	548	591	635	657	35,08 ± 0,70
Knee height (sitting)	knh	458	480	519	556	576	30,64 ± 0,81
Lower leg length with foot	usfl	378	396	433	471	488	25,53 ± 1,01

3.3 Values male

Name of measurment	**Abbr.**	**P 1**	**P 5**	**P 50**	**P 95**	**P 99**	**% from kph**
Body height	kph	1651	1696	1802	1911	1988	100,00
Eye height	auh	1534	1578	1678	1781	1855	93,07 ± 0,14
Chin height	kih	1427	1472	1574	1683	1756	87,16 ± 0,82
Acromial shoulder height	sha	1340	1378	1476	1577	1635	81,71 ± 0,68
Elbow height	ebh	998	1022	1106	1192	1231	61,11 ± 1,06
Knee height	kgh	431	446	489	539	557	26,94 ± 1,05
Foot height	fh	60	66	78	91	95	04,15 ± 0,56
Span width of the arms	spwa	1657	1703	1820	1944	2001	100,88 ± 0,68
Head height	koh	200	204	223	239	246	12,26 ± 0,24
Pupil distance	pd	54	58	63	69	72	03,45 ± 0,17
Arm length (additive)	ala	729	743	795	859	900	44,26 ± 0,57
Upper arm length	oal	299	309	339	370	390	18,63 ± 0,63
Forearm length	ual	231	245	264	289	299	14,55 ± 0,57
Hand length	hdl	172	181	194	210	217	10,71 ± 0,29
Middle finger length	mfl	72	76	82	91	96	04,54 ± 0,20
Thumb length	dlf	61	64	73	86	91	04,00 ± 0,40
Projective thigh length	oslp	302	318	365	410	435	19,69 ± 1,58
Lower leg length	usl	345	446	489	539	557	25,63 ± 3,65
Foot length	fl	239	248	267	290	303	14,77 ± 0,35
Head width	kob	142	146	154	164	170	08,58 ± 0,03
Shoulder width (bideltoidale)	sbh	407	420	455	497	514	25,17 ± 0,68

Elbow width	ebb	62	66	71	77	79	03,90 ± 0,14
Direct hand width	hdb	77	80	87	95	97	04,80 ± 0,15
hand width with thumb	hdbd	94	98	108	120	126	05,94 ± 0,29
Hip width	hueb	302	308	337	372	387	18,66 ± 0,65
Knee width	kb	82	88	96	105	109	05,24 ± 0,26
Fundament-head length	stl	871	885	943	1000	1024	52,40 ± 0,29
Eye height (sitting)	auhs	751	765	821	872	894	45,45 ± 0,26
Acromial shoulder height (sitting)	shas	553	567	619	667	687	34,05 ± 0,74
Elbow height (sitting)	ebhs	185	203	247	296	309	13,09 ± 2,14
Fundament knee length	gkl	556	577	618	666	689	34,21 ± 0,59
Knee height (sitting)	knh	498	515	556	606	627	30,77 ± 0,77
Lower leg length with foot	usfl	412	436	476	519	541	26,06 ± 1,10

[1]A. Freisinger and K. Pfeiffer. Glossary of human kinematics. *IPA internal report*, 2004

Glossar for human anatomy

Greek / Latin	English	German
	Directions of cutting	**Schnittrichtung**
transversal = horizontal	• at right angles to the long axis of the body • from left to right, parallel to the floor	• waagrecht (parallel) zur ebenen Bodenfläche • von links nach rechts
Sagittal = ventrodorsal = anterior – posterior	• in a plane perpendicular to the frontal plane • from behind to the front	• waagrecht (parallel) zur ebenen Bodenfläche • von vorne nach hinten
Longitudinal = vertical	• parallel to the long axis of the body or organ • at right angles to the floor, from up to below	• senkrecht zur ebenen Bodenfläche
	Planes (Sections)	**Ebenen (Schnitte)**
transversal = axial	• parallel to the floor • devide the room into above and below • spread out by sagittal and transversal axes	• waagrecht zur Bodenfläche • teilen den Raum in oben und unten • von sagittaler und transversaler Achse aufgespannt
Sagittal	• at right angles to the floor • devide the room into left and right • spread out by sagittal and longitudinal axes	• senkrecht zur Bodenfläche • teilen den Raum in links und rechts • von sagittaler und vertikaler Achse aufgespannt
Frontal	• at right angles to the floor • devide the room into front and behind • spread out by transversal and longitudinal axes	• senkrecht zur Bodenfläche • teilen den Raum in vorne und hinten • von transversaler und vertikaler Achse aufgespannt
median (sagittal)	Special case: One of the sagittal planes, which devides the human body in two (nearly) symmetrical hafts	Sonderfall: Diejenige Sagittalebene, die den Menschen in zwei (fast) symmetrische Hälften teilt
	General terms of Direction and Position	**Allgemeine Richtungs- und Lagebezeichnung**
• supra- • infra-	• above • below	• oben, oberhalb, über • unten, unterhalb, unter
• anterior, -ius • posterior, -ius	• in front (of) • behind (to the back)	• vorne • hinten
• internus, -a, -um • externus, -a, -um	• within • out of, external	• innen gelegen • außen gelegen
• dexter, -tra, -trum • sinister, -tra, -trum	• right • left	• rechts • links
• cranialis, -e • caudalis, -e	• towards to head • towards the tail	• schädelwärts • steißwärts
• ventralis, -e • dorsalis, -e	• towards the belly • towards the back	• bauchwärts • rückenwärts
• medianus, -a, -um	• in the middle	• in der Medianebene, mittelständig
• medialis, -e • laterialis, -e	• towards the middle of the body • towards the side of the body	• auf die Medianebene zu, mittelwärts • von der Medianebene weg, seitlich
• inter- • intra	• (in) between • within	• zwischen • innerhalb

	Designations for Directions and Positions of the Extremities	**Lagebezeichnungen an den Extremitäten**
• distalis, -e • proximalis, -e	• towards the free end of the extremity • towards the root of the extremity	• rumpffern • rumpfnah
• palmaris, -e • dorsalis, -e	• towards the palm of the hand • towards the back of the hand	• zur Handinnenfläche hin • zum Handrücken hin
• radialis, -e • ulnaris, -e	• on the radial side (radius = spoke of the wheel) • on the ulnar side (ulna = larger bone of the forearm)	• zur Speiche hin • zur Elle hin
• plantaris, -e • dorsalis, -e	• towards the sole of the foot • towards the upper surface of the foot	• zur Fußsohle hin • zum Fußrücken hin
• tibialis, e • fibularis, -e (= peronaeus, -a, -um)	• on the side of the shinbone • on tne side of the calfbone	• zum Schienbein hin • zum Wadenbein hin
	Action of Muscles and Joints	**Funktionen von Muskeln und Gelenken**
• Extension • Flexion	• straigthen an extremity along a (main) axis • Bend an extremity along a (main) axis	• Strecken (Dehnen) • Beugen (Bücken / Neigen)
• Abduction • Adduction	• movement to the side, away from the body / main axis • movement towards the centre of the body / main axis	• Bewegung vom Körper (Hauptachse) weg • Bewegung vom Körper (Hauptachse) hin
• Pronation • Supination	• movement of the forearm to the inner side (e.g. to cut bread) • movement of the forearm to the outer side (e.g. to hold a dish)	• Wendebewegung des Unterarmes nach innen (Brot schneiden) • Wendebewegung des Unterarmes nach außen (Suppe löffeln)
• Eversion • Inversion	• rotation of the feet to the side along an axis from the tip to the heel • rotation of the feet to the centre along an axis from the tip to the heel	• Auswärtskantung der Füße (entspricht Pronation) • Einwärtskantung der Füße (entspricht Supination)
• Anteversion • Retroversion	• move an extremity straight forward from the frontal plane • move an extremity straight backward to the frontal plane	• Vorheben von Arm / Bein (An- / Abwinkeln nach vorne) • Rückführung von Arm / Bein (An- / Abwinkeln nach hinten)
• Rotation	• circular movement round an axis - inwards: movement from neutral position to front and centre - outwards: counter movement (to front and side)	• Drehen / Kreisen um eine Achse - Innenrotation: Bewegung aus der anatomischen Normalstellung erst nach vorne and dann einwärts - Außenrotation: entsprechende Gegenbewegung
• Circumduction	• circular movement (of the thumb)	• Kreisförmiges Herumführen (des Daumens)
• Elevation	• Lifting up an extremity over 90° (to vertical main axis	• Hochheben einer Extremität über 90°! (zur vertikalen Hauptachse)

	Main parts and regions of the body	**Haupteile und wichtige Regionen des Körpers**
• *Caput*	• *head*	• *Kopf*
• Facies	• face	• Gesicht
• Collum / Cervix	• neck	• Hals
• Nucha	• nape	• Nacken
• *Truncus*	• *trunk*	• *Rumpf / Stamm*
• Columna vertebralis	• vertebral column / spine	• Wirbelsäule
• Thorax / Pectus	• thorax / chest / rib cage	• Brustkorb
• Dorsum	• back	• Rücken
• Abdomen / Venter	• abdomen / belly	• Bauch
• Pelvis	• Pelvis	• Becken
• *Membrum superius*	• *Upper extremity / upper limb /arm*	• *Obere Extremität / Arm*
• Braichium	• upper arm	• Oberarm
• Cubitus	• elbow	• Ellenbogen
• Antebrachium	• forearm	• Unterarm
• Manus	• hand	• Hand
• *Membrum inferius*	• *Lower extremity / lower limb/ leg*	• *Untere Extremität / Bein*
• Femur	• thigh (bone)	• Oberschenkel (-knochen)
• Genu	• knee (please note as well: Genu = bend)	• Knie (aber auch: Genu = Biegung)
• Crus	• (lower) leg	• (Unterschenkel
• Pes	• foot	• Fuß
• Akra	• Sticking –out parts of the body, acra (e.g. nosem outer ear, fingertips)	• Vom Körper abstehende Teile, Akren (z.B Nase, Ohrmuschel, Fingerspitzen)
	Bones	**Knochen**
	Basic Terms / Points for Orientation	*Grundbegriffe / Orientierungspunkte*
• Articulatio	• joint	• Gelenk
• Circumferentia	• cirumferent articular surface	• herumführende Gelenkfläche
• Condylus	• condyle	• Gelenkfortsatz
• Epiphysis	• Epiphysis (of a joint) • Pineal gland	• Gelenkfortsatz (langer Röhrenknochen) [im deutschen Gebrauch: Epiphyse (wörtlich: Zuwuchs, Ansatz)] • Zirbeldrüse = Corpus pineale
	Upper Extremity *Bones and Cartilages*	*Obere Extremität (Arm)* *Knochen und Knorpel*
• Humerus	• humerus / [upper arm bone]	• Oberarmknochen
• Rdius	• Radius	• Speiche
• Ulna	• Ulna	• Elle
• Olecranon	• elbow / olecranon	• Ellenbogen(höcker)
• Manus	• Hand	• Hand
• Os scaphoideum	• scaphoid bone	• Kahnbein der Hand
• Os triquetrum	• triquetral bone	• Dreiecksbein
• Os pisiforme	• pisifrom bone	• Erbsenbein
• Os trapezium	• trapezium bone	• großes Vieleckbein
• Os trapezoideum	• trapezoid bone	• kleines Vieleckbein
• Os capitatum	• capitate bone	• Kopfbein
• Os hamatum	• hamate bone	• Hakenbein
• Ossa sesamoidea	• sesamoid bone	• Sesambein

	Lower Extremity (leg) *Bones and Cartilages*	*Untere Extremitäten (Bein)* *Knochen und Knorpel*
• Femur	• thigh bone / femur	• Oberschenkelknochen
• Fibula	• calfbone / fibula	• Wadenbein
• Patella	• kneecap	• Kniescheibe
• Tibia	• shin(bone) / tibia	• Schienbein
• Pes	• foot	• Fuß
• Talus	• anklebone	• Sprungbein
• Calcaneus	• heel bone	• Fersenbein
• Os naviculare	• navicular bone	• Kahnbein des Fußes
• Os cuneiforme mediale	• medial cubeiform bone	• inneres Keilbein
• Os cuneiforme intermedium	• intermediate cuneiform bone	• mittleres Keilbein
• OS cuneiforme laterale	• Lateral cuneiform bone	• Äußeres Keilbein
• Os cuboideum	• Cuboid bone	• Würfelbein
	Vertebral column	*Wirbelsäule*
• Vertebra	• vertebra	• Wirbel
• Vertebrae cervicales	• vertebra of the neck	• Halswirbel
• Vertebrae thoracicae	• vertebra of the chest	• Brustwirbel
• Vertebrae lumbales	• vertebra of the loin	• Lendenwirbel
• Vertebra cervicalis I = Atlas	• first vertebra (of the neck)	• Träger (erster Halswirbel)
• Vertebra cervicalis II = Axis	• second vertebra (of the neck)	• Dreher (zweiter Halswirbel)

REAL-TIME 3D HAND AND ARM TRACKING FROM MONOCULAR COLOR IMAGES

Joachim Schmidt, Jannik Fritsch, Gerhard Sagerer

Applied Computer Science, Faculty of Technology, Bielefeld University
33594 Bielefeld, Germany
{jschmidt, jannik, sagerer}@techfak.uni-bielefeld.de

ABSTRACT

This paper presents a system for tracking a human in 3D based on a videostream from a single camera. The algorithm integrating various image cues in a probabilistic framework can be adapted to the available computational resources. Through using color information for modeling clothing as well as skin color for hands and head, an improved tracking performance is realized. The tracking uses joint angle constraints derived from human kinematics for the prediction of the body configuration but contains no motion priors to enable the tracking of arbitrary motions. With an appropriate configuration, the algorithm runs on a laptop in real-time and tracks the upper body of a human and one arm. This makes it suitable for the recognition of gestures on a mobile robot enabling natural human-robot interaction.

1. INTRODUCTION

A large research area within computer vision is concerned with tracking humans. A rather coarse model of the human body is usually adopted in surveillance applications or for recognizing large-scale activities. Obviously, the recognition of small-scale activities that are related to movements of individual body parts requires a finer human body model. Especially the recognition of hand gestures attracts much interest as humans use their hands frequently, e.g., to communicate information. Many vision-based approaches for gesture recognition have used 2D image data to track human hands. However, gestures are performed in three dimensions and for their correct recognition this third dimension can be crucial. For example, imagine a human pointing to one out of many objects on a table. Here, the object that the human pointed at may be difficult to identify without 3D information about the deictic gesture. Additionally, the fact that the gesture perceived was a pointing gesture may not be detectable without knowledge of the arm configuration.

Our goal is to realize a robot companion that has natural interaction capabilities and can recognize typical gestures occurring in human-human interaction. In order to obtain images of the humans interacting with it, our mobile robot BIRON [1] is equipped with a pan-tilt camera. As BIRON needs to be able to track human hands for recognizing gestures, our aim is to develop a vision-based approach that allows us to track the upper body part of a human in 3D.

As we intend to track human-robot interactions in real-time on a mobile robot, only limited computational power is available. This limitation and the need to enable tracking of humans with arbitrary clothing rules out approaches using stereo vision. Such approaches require image disparities to successfully carry out the computationally heavy depth calculation. Instead, we use a 3D model of the upper human body and map this model to the 2D image data. For the limbs we use several cues derived from color images and skin color information is used as a cue for detecting hands and faces. Tracking the 3D model over time is realized using a probabilistic algorithm. Special emphasis is placed on the scalability of the framework to enable the real-time calculation of a 3D body configuration. First results demonstrate the suitability of our approach for tracking hand gestures using a standard laptop.

The paper is organized as follows: Section 2 gives an overview of related work. The individual image cues and the probabilistic tracking method is described in Section 3. The configuration of the overall framework to enable real-time tracking and the results obtained are the topic of Section 4. The paper concludes with a summary in Section 5.

2. RELATED WORK

One of the first applications to track a human body in real-time using a single camera is the PFINDER system [2]. The applied body model is rather coarse and tracking provides only the 2D shape with information about the position of head, hands, and feet.

Tracking of a human in 3D with limited computational resources on a mobile robot was already described in 1996 by Kortenkamp et al. [3]. However, this approach used depth information from a stereo camera to track the 3D body model. In more recent approaches, the skin color is used as an additional cue to get the 3D hand position (see, e.g., [4]).

As the depth calculation from stereo images comes at a high computational cost and relies on the presence of image

disparities, the use of a monocular approach for estimating the 3D configuration is a promising alternative. One such approach for tracking a detailed 3D human body model was proposed by Sidenbladh [5]. It is based on a variety of gray-level image cues that are combined using a particle filtering algorithm for tracking the motions of a human. To cope with the huge search space, motion priors are used to predict the changing 3D body configuration. Following Sidenbladh's work, Sminchisescu used a more precise modeling of the 3D body model [6], but the computational time required prohibits its use for real-time tracking.

3. TRACKING A 3D HUMAN BODY MODEL

In the following we describe our framework for tracking a 3D human body model. Our architecture is based on Sidenbladh's approach [5]. The pose of an articulated 3D body model is rated to comply with the current image of the person tracked depending on several image cues. Merging the likelihood of the cues gives an estimation for each pose. The search for the best model can be seen as an optimization process of finding the global maximum in the high-dimensional feature space of the joint angles. According to a probabilistic approach using particle filtering, the systems finds a good-fitting pose of the body model.

3.1. Modeling the Appearance of Humans

We use an articulated body model consisting of cylinders with ellipsoid cross sections to represent the appearance of a human body. Joint angle limits model the physical constraints of the human body. A typical model of an upper body contains up to 23 degrees of freedom, a coarse full body model up to 34 degrees of freedom [5].

To reduce the numerical complexity for the following calculations, the 3D body model is backprojected into the image plane using a pinhole-camera model. This yields an approximate 2D representation of the 3D body model consisting of a 2D polygon with image coordinates for each limb. The estimation of a single pose is done by calculating likelihoods for each limb of the body model with up to four image cues:

1) Similar to [5], the **Edge Cue** uses the first partial derivatives that are sensitive to strong changes in contrast. For recognizing human body parts mainly the presence of edges is important, not their magnitude. All partial derivative images are, therefore, scaled with a nonlinear localized normalization function to smooth low energetic edges stemming from textured backgrounds and emphasize stronger ones. The Edge Cue provides an accurate match for the position of arm limbs by comparing the direction of the edges in the camera image with the expected orientation of the limb at a number of feature points.

2) The **Ridge Cue** uses the normalized second partial derivatives (see Fig. 1) to find elongated structures of a specified thickness in the image, also similar to [5]. It suppresses point-like edge features by looking at a number of feature points for edges parallel to the expected limb orientation and missing edges in perpendicular direction. The Ridge Cue gives a coarser estimate of the limb position than the edge cue, but it produces less false maxima. The Ridge Cue depends on the size of the limbs in the image. This means that it will only provide appropriate results if the observed limb is in a particular distance to the camera. But as the distance between the person and the camera is not fixed, the size of the limbs observed in the picture may vary. Creating a Gaussian image pyramid allows the Ridge Cue to adapt to different distances by selecting the correct resolution level in the pyramid based on the current distance of the model.

Fig. 1. First partial derivative in Y direction

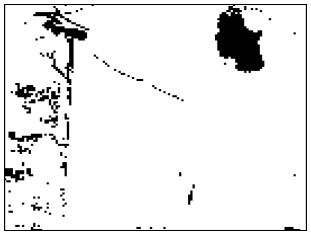


Fig. 2. Result of skin color segmentation

3) The **Mean Color Cue** substitutes Sidenbladh's Flow Cue, as the latter has disadvantages in terms of its robustness. Sidenbladh's Flow Cue samples pixels at corresponding positions on the limbs in successive grey level images. Repeating pixel patterns on a limb can be found, but changes in the appearance of a limb (e.g., shadowing, folds forming in clothing while moving) pose problems. Furthermore, as the Flow Cue only relates to the estimated pose of the body in the previous time step, small tracking errors are accumulated and the limb tends to drift.

In contrast, the Mean Color Cue deals with these problems by using multiple mean color values for each limb (see Fig. 3) that are adapted over time. Each mean color value is calculated by averaging over the color value of N pixels sampled from the 2D polygon of the backprojected limb. Sampling is performed for each so-called blob using a 2D Gaussian distribution around a center position. The number of blobs and their center positions are chosen based on the limb type.

For likelihood calculation, the actual mean color value for each blob is compared to its slowly adapted mean value. Updating is realized using the mean color values of the blobs belonging to the estimated correct pose. To deal with tracking errors and changes in the appearance of the human, we use a sliding adaption process and update the mean color only with a small fraction.

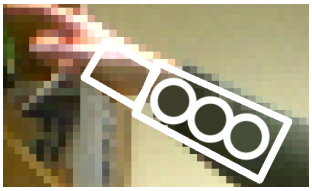


Fig. 3. Position of blobs for mean color cue

Overall, the Mean Color Cue gives a very reliable estimation for the limb position, as color is a strong characteristic. It can deal with shadowing and changes in the appearance. Additionally, it is less sensitive with respect to minor disturbances and is therefore more robust than the original Flow Cue.

4) The **Skin Color Cue** uses a skin color segmentation [7] to gain information about the position of the hand and the head. The skin color model is initialized using a Viola-Jones face detector (see [8]) and provides a binary skin image (see Fig. 2). Similar to the Mean Color Cue, sample pixels are selected from Gaussian distributed blobs. The likelihood of a limb is calculated as the ratio of pixels being classified as skin and non-skin.

The current implementation does not take advantage of the probabilistic modeling of skin (see [7] for details). Nevertheless, this cue is very useful in environments without strong lighting variations. If lighting conditions change constantly, the adaptation of the skin color model can be used at the cost of an additional computational load.

3.2. Combination of Likelihoods

The Edge, Ridge, and Mean Color Cues generate a number of likelihoods depending on the number of feature points or blobs. These likelihoods have to be combined to obtain a likelihood for the alignment of a single limb. The Skin Color Cue provides just one likelihood per limb (hand or face). Subsequently, the limb likelihoods are used to compute a single likelihood for each pose of the body model.

Sidenbladh assumes that the cues and limbs are independent, hence she calculates the overall likelihood for a single particle proportional to the product over all feature points, all image cues, and all limbs. On the basis of this multiplication, a single poorly rated feature point can drag down the whole likelihood for a pose. This, in turn, leads to a peaked probability distribution that is harder to evaluate as the correctly estimated pose may be represented by only a small percentage of all particles.

In our approach, in contrast, all likelihoods for a single cue on each limb are averaged, smoothing the effect of outliers. The following multiplication over all cues and all limbs is therefore less problematic. A further advantage is the higher flexibility when changing the number of feature points for a single cue, as this does not change the number of likelihoods that are multiplied to obtain the pose likelihood.

3.3. Probabilistic Tracking

The challenge of tracking a human can be formulated as finding a maximum in the high-dimensional joint angle space. The distribution is represented by a set of particles (see Fig. 4) that are propagated using particle filtering.

Every particle represents a different pose of the human body. The propagation over time is accomplished by adding Gaussian-shaped noise to all joint angles. Predefined motion primitives or an overall dynamic model are not used to enable the tracking of arbitrary motions.

The estimation of the best fitting model is realized by averaging over the distribution of all particles with respect to their likelihood. The outcome of this is one single pose (see Fig. 5), from which the 3D trajectory data can be extracted (example in Fig. 6) and which also serves as reference model for updating the mean color values.

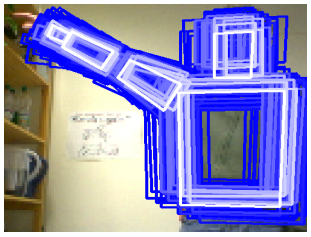


Fig. 4. 100 best particles, colored by their likelihood

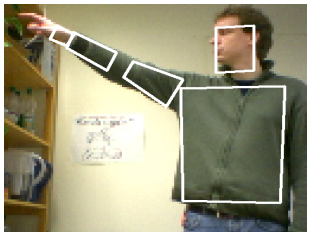


Fig. 5. Estimation of the current pose

With linear scaling of the computational complexity corresponding to the number of particles, the robustness and quality of the tracking can easily be adopted to a specific hardware or real-time requirements.

4. CONFIGURATION AND RESULTS FOR REAL-TIME TRACKING

For the intended application of pointing gesture recognition on a mobile robot, we track only the upper body with the right arm assuming deictic gestures are performed with the right hand. In our current model the hand is fixed to the lower arm and the head is fixed to the torso leading to a model with 10 degrees of freedom. This reduction of the model flexibility is acceptable for tracking a human that is oriented roughly towards the camera and performs communicative gestures.

For the implementation on BIRON we could have used the position of a person relative to the robot from our multimodal person tracking system [8], but experiments have shown that it is sufficient to track the skin-colored head and link the torso to the head. Consequently, no cues are

calculated for the torso to save computing time. The Skin Color Cue is used for head and hand with 10 feature points each, while the two arm limbs are tracked by combining the Edge, Ridge, and Mean Color Cue. For calculating the Mean Color Cue we use three blobs per arm limb, each with 10 feature points. The Edge Cue uses 20 randomly sampled points per limb, the Ridge Cue 30. These parameters have been found to be an acceptable tradeoff between robustness and computational load.

Initialization of the overall torso position is done using our implementation of the Viola-Jones face detector (see [8]). Currently no automatic initialization of the arm and hand is implemented, but this could be done using information from skin segmentation.

Note that in the current implementation no occlusion check is performed to guarantee a fixed processing time. As no cue for the torso is calculated, this is only problematic if the hand occludes the head or points in the direction of the camera. In such situations correct tracking may temporarily fail, but usually the probabilistic nature allows the algorithm to 'snap in' again on the arm when the occlusion ends.

The algorithm was tested on an IBM Thinkpad X31 (Pentium M, 1.4 GHz) with an Apple ISight FireWire camera. Using 220 particles and images with 160×120 pixels gives the current 3D hand position at a framerate of 7.5 Hz. Figure 6 shows the 3D hand trajectory of a human performing two pointing gestures with his right arm (for example movies see [9]).

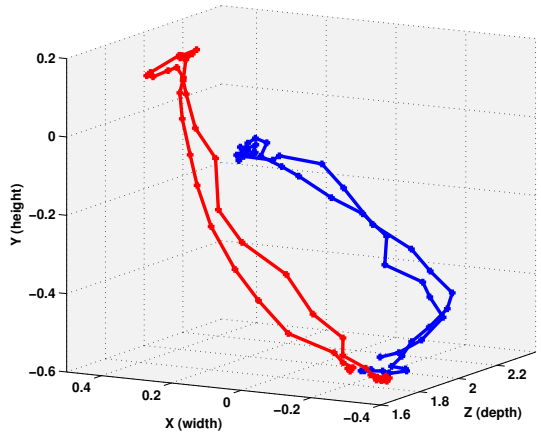


Fig. 6. Tracked 3D hand trajectory

With the configuration described above, tracking of hand gestures has been successfully performed in environments with a complex background. The rather coarse estimate of the 3D hand position due to the small image resolution is sufficient for a rough estimation of pointing gestures. If no large self-occlusions occur, our framework is able to successfully track a human on a mobile robot using a standard camera and a laptop.

5. SUMMARY

We presented a probabilistic framework using a variety of image cues to track a human in 3D based on 2D image data. Robust tracking is achieved by the use of adaptive color image cues for clothing and skin in addition to ridge and edge cues. The framework is flexible as it can be configured for real-time performance with limited computational resources by tracking only a subset of the human body. First results reaching a framerate of 7.5 Hz demonstrate its suitability for the recognition of pointing gestures on a mobile robot.

6. REFERENCES

[1] A. Haasch, S. Hohenner, S. Hüwel, M. Kleinehagenbrock, S. Lang, I. Toptsis, G. A. Fink, J. Fritsch, B. Wrede, and G. Sagerer, "BIRON – The Bielefeld Robot Companion," in *Int. Workshop on Advances in Service Robotics*, Stuttgart, Germany, 2004, pp. 27–32.

[2] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Trans. Pattern Analysis and Machine Vision*, vol. 19, no. 7, pp. 780–785, 1997.

[3] E. Huber D. Kortenkamp and R. P. Bonasso, "Recognizing and interpreting gestures on a mobile robot," in *Proc. Nat. Conf. on Artificial Intelligence*, 1996, pp. 915–921.

[4] E. Seemann K. Nickel and R. Stiefelhagen, "3D-Tracking of Heads and Hands for Pointing Gesture Recognition in a Human-Robot Interaction Scenario," in *Int. Conf. on Face and Gesture Recognition*, Seoul, Korea, 2004.

[5] H. Sidenbladh, *Probabilistic Tracking and Reconstruction of 3D Human Motion in Monocular Video Sequences*, Ph.D. thesis, KTH Sweden, 2001.

[6] C. Sminchisescu and B. Triggs, "Mapping minima and transitions of visual models," *International Journal of Computer Vision*, vol. 61, no. 1, 2005.

[7] J. Fritsch, *Vision-based Recognition of Gestures with Context*, Ph.D. thesis, Bielefeld University, 2003.

[8] J. Fritsch, M. Kleinehagenbrock, S. Lang, T. Plötz, G. A. Fink, and G. Sagerer, "Multi-modal anchoring for human-robot-interaction," *Robotics and Autonomous Systems*, vol. 43, no. 2–3, pp. 133–147, 2003.

[9] J. Schmidt, "Example movies showing 3D human body tracking," `http://aipc1.techfak.uni-bielefeld.de/~jschmidt/`.