AUTOMATIC INITIALIZATION FOR BODY TRACKING
Using Appearance to Learn a Model for Tracking Human Upper Body Motions

Joachim Schmidt
Bielefeld University, Technical Faculty, Applied Computer Science
P. O. Box 100131, D-33501 Bielefeld, Germany

Modesto Castrillón-Santana
Institute of Intelligent Systems and Numerical Applications in Engineering
Edificio Central del Parque Tecnológico, Campus de Tafira, University of Las Palmas de Gran Canaria - Spain

Keywords: Human robot interaction, face detection, model acquisition, automatic initialization, human body tracking.

Abstract: Social robots require the ability to communicate and recognize the intention of a human interaction partner. Humans commonly make use of gestures for interactive purposes. For a social robot, recognition of gestures is therefore a necessary skill. As a common intermediate step, the pose of an individual is tracked over time making use of a body model. The acquisition of a suitable body model, i.e. self-starting the tracker, however, is a complex and challenging task. This paper presents an approach to facilitate the acquisition of the body model during interaction. Taking advantage of a robust face detection algorithm provides the opportunity for automatic and markerless acquisition of a 3D body model using a monocular color camera. For the given human robot interaction scenario, a prototype has been developed for a single user configuration. It provides automatic initialization and failure recovery of a 3D body tracker based on head and hand detection information, delivering promising results.

1 INTRODUCTION

As stated by McNeill gestures play an important role in human communication (McNeill, 1992). According to this, a social robot requires the abilities to localize, track and interpret human behavior during an interaction session. Pursuing this challenge is an active field in Computer Vision research, due to the restrictions imposed by approaches developed so far for this purpose.

A simple 3D body model based on joints and their motion is sufficient for humans to perform action recognition (Johansson, 1973). 3D body acquisition and tracking provides the data source necessary to accomplish such a task, offering a great domain of applications: Surveillance, Activity Recognition, Human Computer/Robot Interaction, Mobile Robotics, etc.

The task of tracking a human body in monocular images is commonly carried out by fitting articulated kinematic models representing the appearance of a person to mono- or multiocular images (Sidenbladh et al., 2000; Sigal et al., 2004; Sminchisescu and Triggs, 2005). The surveys of Gavrila (Gavrila, 1999), and Moeslund and Granum (Moeslund and Granum, 2001) provide a good overview of the topic of model based 3D body tracking. In addition to a kinematic body model describing the appearance of the body, bringing in prior knowledge about familiar body configurations (Brox et al., 2006) can help to constrain the search process, preventing the production of unrealistic pose estimates. Difficulties in recognizing ambiguous poses, as is common in monocular tracking, can be overcome when 3D information is available, e.g., acquired from time-of-flight sensors or stereo camera systems, as shown by (S. Knoop, 2006). For most of the above approaches, however, the question of initialization stays open or is subject to manual or semi-automatic procedures. Using body tracking in human robot interaction often comes with strong restrictions, e.g., the number and type of cameras available or the gesture repertoire to be observed. Tracking systems incorporating automatic or semi-automatic initialization often make use of learned appearance models (Ramanan and Forsyth, 2003), rely on stereotyped poses (Urtasun et al., 2005) or on combining a repertoire of learned pose estimates and visual appearance (Sigal and Black, 2006a; Tay-

Schmidt J. and Castrillón-Santana M. (2008).
AUTOMATIC INITIALIZATION FOR BODY TRACKING - Using Appearance to Learn a Model for Tracking Human Upper Body Motions.
In *Proceedings of the Third International Conference on Computer Vision Theory and Applications*, pages 535-542
DOI: 10.5220/0001071005350542
Copyright © SciTePress

cher et al., 2006; Bissacco et al., 2007). Initialization can also be formulated as the problem of pose estimation or object reconstruction from a single image using strong models (Lee and Cohen, 2004).

Summarizing, recent literature has described different approaches focused on tracking people. However, there is still a gap between tracking algorithms and systems working in the real world, mainly due to the fact that for most tracking approaches the challenges of automatic initialization and error recovery are not addressed. In this paper, we present a system within a typical human robot interaction scenario, i.e., where an individual is communicating with an artificial actor. The basic idea guiding the design of our system is to integrate robust face and hand detection results into a model representation that can be used for automatic initialization and failure recovery of a pose tracking algorithm.

Section 2 gives an overview about the functionality of the approach describing the system modules in detail. The experimental results and conclusions are presented respectively in Sections 3 and 4.

2 SYSTEM

2.1 Overview

In this section we present our approach of a self-starting 3D body model tracker in a human robot interaction scenario.

There are a number of features that should be considered in order to make a system flexible enough to operate within this context: 1) The person and its body dimensions must not necessarily be known a priori, but the distance of the human to the robot has to be adequate for interaction. 2) Images are acquired using a single monocular color camera as the system is intended to be used on a mobile robot without employing further sensors. 3) During system design, we also avoided specific background models to allow the tracking to be independent from the appearance of the observed scene. To further allow a moving camera, image background substraction based techniques like motion history images are avoided as well.

For the proposed initialization procedure, some restrictions, well suited for human robot interaction, can be derived from the given scenario: 1) The person is trying to communicate, therefore his intention is to cooperate with the system. 2) The upper body, including the head, torso and both hands, is visible and no large self occlusion occurs. 3) The person is standing in an upright position, facing the camera and having the arms outstretched.

Algorithm 1: Algorithmic processing overview.

estimate initial pose from single image:
if *face is detected* **then**
- extract skin and shirt samples
- update color models for skin and shirt
- create segmented images
- detect hands using skin and shirt color segmentation
- **generate initial density estimating the correct pose:**
- **if** *hands are detected* **then**
  - find probable model poses based on the distances of the hands and the face, use 5 DOF for model
- **else**
  - body model facing the camera, arms hanging down, use 3 DOF for model

track human body:
wait until initial density is provided
if *tracking for the first time* **then**
- **start tracking:** use density as prior

else
- **keep tracking:**
- **if** *initial density is provided* **then**
  - use density as recovery component,
  - use updated skin and shirt color models for tracking
- **else**
  - track relying on current model

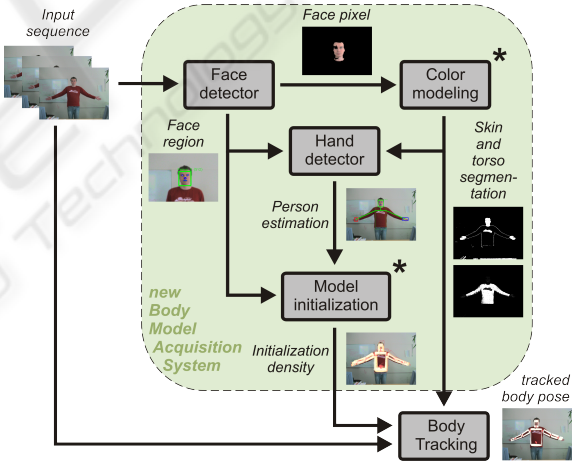


Figure 1: System overview. Faces are detected, shirt and skin color are learned, hand and face positions are used to estimate a density for initializing the tracking and for error recovery. Modules denoted by *. have been used in the existing body model tracking framework, but had to be manually initialized.

Succinctly the process, depicted in Figure 1 and briefly outlined in Algorithm 1, applies first face detection. A face has to be detected for estimating an initial pose density and, starting the tracking. Its color is modeled and based on this, both hands are located. Face and hands information, if detected, are utilized to estimate the likelihood for an initial body model location. Once initialized, the body tracker performs continuously considering face and hands information

whenever available. Future face and hand features allow the tracking system to have an additional validation control useful to recover the tracking from failures.

The different system modules are described in the following subsections.

2.2 Face and Facial Element Detection

Several approaches have recently appeared presenting reliable face detection in real time (Schneiderman and Kanade, 2000; Viola and Jones, 2004), making face detection less environment-dependent. However, cue combination can provide greater robustness and higher processing speed, particularly for live video stream processing (Castrillón Santana et al., 2007), outperforming single cue based detectors such as (Viola and Jones, 2004).

The face detection system (Castrillón Santana et al., 2007), integrates among other cues, different classifiers based on the general object detection framework by Viola and Jones (Viola and Jones, 2004), skin color, multilevel tracking, etc. The chosen detection system provides not only face detection but also facial feature location information in many situations. Extending a face detector with inner feature detection (eyes, nose and mouth) also reduces the number of false alarms. As more restrictions are imposed, it is less likely that all the detectors are activated simultaneously with false alarms, thereby minimizing the influence of such errors.

Figure 2: Normalized face sample and likely locations for nose and mouth positions after normalization.

Positive samples were obtained by annotating manually the eye, nose and the mouth location in 7000 facial images taken randomly from the internet. The images were later normalized by means of eyes information to 59×65 pixels, see Figure 2(left). Five different detectors were computed: (1,2) Left and right eye (18×12 pixels), (3) eye pair (22×5), (4) nose, and (5) mouth (22×15).

The facial element detection procedure is only applied in those areas which bear evidence of containing a face. This is true for regions in the current frame, where a face has been detected, or in areas with detected faces in the previous frame. For video stream processing, given the estimated area for each feature, candidates are searched in those areas not only by means of Viola-Jones' based detectors, but also by sum-of-squared differences (SSD) tracking previous facial elements. Once all the candidates have been obtained, the combination with the highest probability is selected and a likelihood based on the normalized positions for nose and mouth is computed for this combination, see Figure 2.

2.3 Color Modeling

To add further details to the representation of the individual detected, our system learns models of the skin and the shirt color taking into account the previously detected face region. This is done only for a robustly detected face, for which at least three facial features - the face itself, its eyes and the nose or the mouth - have been detected as a trusted result. This consideration is used to reduce the possibility of false detections, i.e. false positives. A color model is then learned or updated (if already created for that individual) only from these trusted faces, reducing the probability of using erroneous face detections.

(a) (b) (c)

Figure 3: (a) Input image, (b) skin color segmentation, (c) shirt color segmentation.

Eye locations allow the estimation of the face container. In a lower position a second container is used to select an area of the user's shirt. Both containers are utilized to model respectively the skin and shirt colors of the user by means of a histogram based color model (Swain and Ballard, 1991). Figure 3 presents the segmentation of the input image, based of these histograms, see Figure 3(b) for skin and Figure 3(c) for shirt color-like areas. It can be observed that the shirt is segmented easily but hands and skin-like areas, are not necessarily clear and compact.

Due to the sensitivity of the histogram based skin color model, we analyzed another color model representation. The mask of the skin blob extracted from the face container determines the skin pixels to be employed as training samples, see Figure 4(a). The skin color of each individual is then learned and further adapted throughout tracking as a mixture of gaussians in RG-color space (Fritsch et al., 2002), see Figure 4.

2.4 Hand Detection

Multiple difficulties are present regarding robust and efficient hand detection in video, mainly due to the

inherent variability of the articulated hand structure, the large domain of gestures, the restriction of real-time performance, varying illumination conditions and complex background clutter. Therefore, different restrictions are commonly considered or even manual initialization is performed for this task.

However, the literature is rich in hand detection approaches that have traditionally been based on skin color segmentation (Storring et al., 2004), due to their reduced processing cost. Recent approaches (Stenger et al., 2004; Kölsch and Turk, 2004), however, have utilized the Viola-Jones' object detection framework (Viola and Jones, 2004) even when hands are not that easy to describe as faces. They are highly deformable objects, so training a single cascade classifier for detecting hands is a complex and arduous task. For that reason, a different classifier for each recognizable gesture has been trained (Stenger et al., 2004), but also a single classifier for a limited set of hands (Kölsch and Turk, 2004).

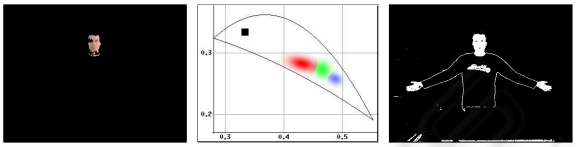
Considering the unrestricted context considered, where the use of multiple detectors would produce an approach not suitable for real time processing, we have chosen the skin color approach for faster processing. However, instead of using a predefined color space definition, the information obtained from the face blob is used, as described above, to estimate the skin color model for that individual, see Figure 4. The skin color model is then employed to locate other skin-like blobs in the image.

As we mentioned above, the approach considers that both hands are visible, no gloves are used, their distance to the face is similar, and that a vertical line falling from the face center would leave each hand on one side. If all those conditions fit, and two well proportionated and coherent skin blobs are located then they are suggested as hands candidates, and provided to the 3D tracker initialization module, see Figure 5.

2.5 Human Body Model Acquisition

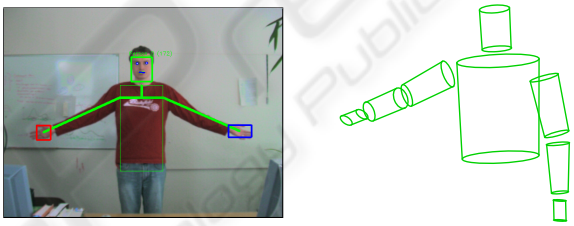
In our approach, the human body is modeled as a kinematic 3D body model, see also (Schmidt et al., 2006), composed of asymmetrically truncated cones connected by joints as depicted in Figure 5. It resembles the kinematic structure of the human body and also incorporates the natural joint angle constraints. In total, a pose of our model can be described by 14 degrees of freedom (DOF) for the joint angles (4 per arm) and the position and orientation in space (3 translational, 3 rotational).

The appearance of the human, in turn, is modeled by the truncated cones. The proportions of the body parts vary for each individual, but still, its characteristics can be described quite well using a standardized model, e.g., (Humanoid Animation Working Group, 2007) and adjusting the actual size of the limbs and the kinematic configuration in proportion to the height of the person. In this approach, we use a standard body model and vary the overall size in the image only, the size of the limbs and the kinematic structure is kept constant.



(a) (b) (c)

Figure 4: (a) Skin color training pixels produced from the face detection, (b) example skin locus and learned mixture of gaussians in RG-color space, (c) resulting segmentation.



(a) (b)

Figure 5: (a) Detected head and hands location, (b) articulated 3D model for the upper body.

Estimating the correct pose is even more difficult on a single frame than on a sequence where also temporal dependencies could be used. Restrictions on the scenario and the expected poses help to simplify the task. For initialization, we assume the person to be facing the robot with the arms outstretched. As a consequence, the model can be constrained to be oriented towards the camera and the arm limbs to move only in a plane parallel to the image plane, having the elbow relaxed. This allows reducing the number of DOF to be determined during initialization to only 5 parameters: Three for the model position and one for the elevation of each arm. In situations in which no hands can be found, the arms are assumed to be close to the torso, reducing the dimensionality d of the parameter space further to 3.

The 2D distances between the detected face, d_F, and left and right hand features, d_{HL} and d_{HR}, and the corresponding model limbs are converted into likelihoods using the following Gaussian weighting function:

$$p(c) = \exp\left(-\frac{(d_c)^2}{2\sigma_c^2}\right) \quad (1)$$

where the standard deviations σ_c are choosen to cover the maximal observable distance depending on the

image size for each utilized feature $c \in \{F, HL, HR\}$. For a number of different poses, these distances are almost the same, as the positions of the face and the hands in the image will not change drastically when translating the model in the depth direction. The likelihood $p(c)$ that a model pose $\mathbf{x}_t$ at the current time t causes the observation $\mathbf{y}_t$ can be formulated as

$$p(\mathbf{y}_t \mid \mathbf{x}_t) = \prod_{c \in \{F,HL,HR\}} p(c) \tag{2}$$

with $p(c) = 1$ if the feature is not present. In the d-dimensional space $\mathcal{R}^d$ of all possible poses, an initialization particle set $\mathbf{SI}_t = \{\mathbf{s}_t^{(n)}\}_{n=1}^N$ is used to represent the observation density with the associated weights $\{w_t^{(n)}\}_{n=1}^N$ distributed according to $p(\mathbf{x}_t \mid \mathbf{y}_t)$ and w normalized to $\sum_{n=1}^N w_t^{(n)} = 1$.

Using this particle set, we employ a kernel particle filter for searching the pose space for initialization postures agreeing with the results from the face and hand detection. The result is a particle distribution estimating the likelihood density in the reduced parameter space. The distribution indeed shows a good estimation seen from the image plane perspective, but it covers a wider range in the depth direction, as the distance to the camera respective the size of the model can only be determined indirectly.

Up to this point, the problem of fitting the model to three given points in the image can be solved much more easily, e.g., making use of inverse kinematics. But using a multiple-hypothesis approach for both tracking and initialization gives us the advantage of allowing for the initialization procedure to generate more than one possible outcome and also leaves room for further extensions of the presented approach, which will possibly result in higher dimensional statespaces.

2.6 3D Body Model Tracking

For estimating the pose of the human in a new frame, we apply the monocular 3D body model tracking algorithm presented in (Schmidt et al., 2006). The human upper body is tracked using the model depicted in Figure 5 with 14 DOF. To match a given pose of the model with the image data, the 3D body model is backprojected into the image resulting in a 2D polygon representation. For each limb, several cues based on color (skin, mean color) and/or intensity (edge, ridge) are evaluated. To give an example, the edge cue evaluates the intensity gradient at the boundary of the 2D polygon representing the arm limb. The likelihood for a specific pose is obtained by fusing the filter responses for all cues and all limbs and transferring them into likelihoods with a cue-specific weighting function representing the expected characteristics of the cue. Combining multiple cues makes the estimation more robust against local disturbances but typically also results in a high number of false local maxima in the parameter space.

The problem of estimating the correct pose of the model for each image can now be expressed as finding the one local maximum in the high-dimensional parameter space, that fits best the current pose while still obeying all given constraints, e.g., the joint angle limits. This pose estimation for a complicated articulated model from monocular observations is a highly ambiguous task. The resulting posterior likelihoods over human pose space are typically multi-modal with a high number of false local maxima. When evolving over time, new modes often emerge from regions with low probability while existing modes degenerate or even vanish. To track the correct pose of the human, the structure of the high-dimensional probability density has to be efficiently exploited, taking into account the constraints and the dynamics of the model. The utilized kernel particle filter (Schmidt et al., 2006) propagates such multi-modal distributions and provides a probabilistic search for the best matching body configuration.

A major problem of all tracking approaches is the tendency to get stuck in false local maxima. To overcome this drawback, the presented approach adds a recovery component to the existing tracking framework. Recovering from tracking errors is achieved by inserting a fixed percentage α (e.g., 5% - 20%) of particles from the initialization distribution $\mathbf{SI}_t$ into the tracking distribution $\mathbf{ST}_t$:

$$\{\mathbf{ST}_t\}_{n=1}^{\lfloor \alpha \cdot N \rfloor} = \Phi(\mathbf{SI}_t, n) \tag{3}$$

with $\Phi(\mathbf{SI}_t)$ sampling from the distribution according to the weight of the particles. These particles do not necessarily represent the correct pose and in most cases will be neglected during the tracking process. If the tracking gets stuck in a wrong pose, the recovery particles explore the parameter space in a region outside of the current search radius of the tracking process while they are still more directed - and therefore more likely to resemble the correct pose - than randomly distributed particles.

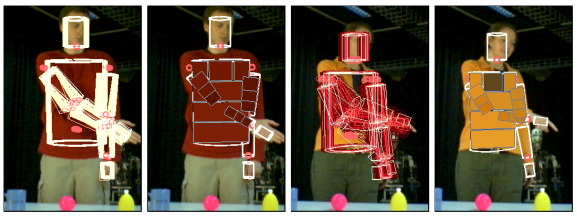
2.7 Integration

All the modules have been implemented as plugins for a graphical plugin shell (Lömker et al., 2006), which provides a framework for image grabbing, display, etc. The plugins are written in C/C++, making some of them use of the OpenCV (Intel, 2006) libraries for image analysis. An integration architecture (Fritsch

and Wrede, 2007) is used for communication between different modules, which allows us to split the system up into multiple instances distributed to different computers in a network or to different cores on the same computer. Currently, the system is divided into four instances: 1) face and hand detection, 2) color modeling, 3) initial distribution generation and 4) body model tracking.

3 EXPERIMENTS

For evaluation, the system has been applied to image sequences of three persons pointing at objects on a table with 836 frames in total.



(a) (b) (c) (d)

Figure 6: Results for subjects B (a-b) and C (c-d) from tracking with learned body model: (a) and (c) show the most likely poses, (b) and (d) the final tracking results.

Given the position of the face in an image, either the height of the person or the distance to the camera can be determined due to the monocular setup. For the experiments we assumed a fixed height for the person, thus only varying the distance during initialization.

Ground truth has been generated by manually annotating the position of the hands and the head. To show the effectiveness of the presented approach both for initialization and recovery from tracking failures, the automatic initialization setup is compared to the manually initialized setup as described in (Schmidt et al., 2006). For the latter, we still need a relatively high number of particles to ensure robust tracking over a longer image sequence, furthermore it is necessary to adjust the measures of the body model for each person accurately. We configure the system in two ways: 1) 1500 particles and 6 meanshift iterations, 2) 500 particles and three meanshift iterations. The automatically initialized system is also set up to use 500 particles, three mean shift iterations but uses a generic body model for all subjects. For each iteration, $\alpha = 5\%$ recovery particles are inserted into the tracking distribution. All setups employ the same cues with identical parameterization.

Figure 6 shows typical tracking results using the automatically acquired body model. For Figure 6(a) and (c) the most likely poses are colord white, less likely poses red. Note the multi-modal distribution in Figure 6(c) due to ambiguous measurements of the pose. Figure 6(b) and (d) show the final tracking result and the learned shirt color.

Table 1: Position error RMSE (root mean squared error) and standard deviation σ in pixel. Comparison of three setups: manual initialization with 1500 particles and with 500 particles, automatic initialization and error recovery with 500 particles (presented approach). The colored mean error values for subject B can also be seen in Figure 7.

		manual initialization				automatic init	
		1500 particles		500 particles		500 particles	
sequence	# pict.	RMSE	σ	RMSE	σ	RMSE	σ
subj A	318	18.73	13.87	52.65	41.31	30.11	26.05
subj B	242	14.52	8.58	32.86	23.26	21.96	22.13
subj C	276	12.04	10.14	72.64	38.30	54.82	57.76

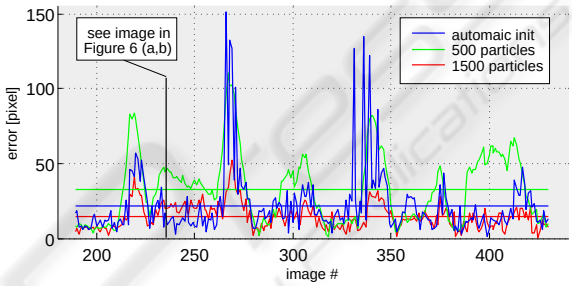


Figure 7: Comparing tracking errors of subject B for manual and automatic initialization setups. Also compare to Table 1. Note the tracking loss around frames 265 and 340 and the recovery afterwards.

Table 1 shows the tracking quality as the differences between the annotated position of the right hand and the model position projected into the image plane as RMSE and standard deviation in pixel for each sequence. For the presented approach, the RMSE stays between 20 – 55 pixel, which is accurate enough for detecting gestures in a human robot interaction scenario (Haasch et al., 2005). In contrast to the existing tracking approach, the standard deviation is much higher for the presented approach, which suggests that the tracking suffers from losses but is often able to recover again as depicted in Figure 7 taking subject B as an example (blue line). Tracking gets lost around frames 265 and 340, but as soon as the situation gets less complicated, the inserted recovery particles are able to guide the tracking towards the correct pose again. Thus, the loss is only temporary and results in a comparably low RMSE as for 1500 particles (red line), but using only a third of the number of particles. Employing an identical parameterization using 500 particles but without the automatic initialization and recovery behaviour leads to an even less accurate tracking (green line) with a 30% higher error in aver-

age. For subject C, the system tends to lose tracking over and over again for sophisticated postures, e.g., the hand pointing directly towards the camera, where also the recovery component does not work resulting in high errors for both approaches. The scaled standard body model does not suit this subject well enough. This is also reflected by the high RMSE of more than 70 pixel. Actually, the tracking was stuck in one position while the person moved, producing varying error values. Robust tracking for this subject is nevertheless possible. The former approach applied a personalized body model and an increased computational effort, yielding accurate results. This clearly shows the limits of the presented approach and calls for automatically adapted model kinematics and limb sizes

The usability of the presented system for the human robot interaction is much higher compared to the former tracking approach. Persons interacting with such a system now have the possibility to repeat unrecognized gestures.

4 CONCLUSIONS AND OUTLOOK

In this paper we have defined an approach to acquire a human body model in the human robot interaction context. In that scenario, it is expected that the human will try to make himself or herself visible. We can therefore make use of a currently available face detector as a starting hint.

Following this idea, we have built a prototype system that makes use of a robust face detection algorithm and some common sense considerations concerning the given task and scenario during a human robot interaction session. The system is able to provide initialization data and also validation data to a tracking framework. This second feature is of great interest for building a robust 3D body tracker. The main features of this prototype are that the process is performed close to real time with a monocular and uncalibrated camera, without the requirement of markers nor the restriction of a static camera. The system is currently performing asynchronously. Face and hand detection perform in real-time while initialization and tracking still require more computational power and provide results in the order of one frame per second without optimized code.

The prototype presented here has tackled a reduced number of sequences, the achieved results are promising. Appearance based initialization of the model and initialization of the body tracking system performs robustly throughout the applied sequences, making the system more usable for human robot interaction. The possibility to validate the tracking and recover it from errors is an important step towards robust system design.

For that reason, in short term the prototype requires a harder experimental evaluation setup considering different scenarios (e.g., sitting at a desk), more sophisticated conditions (e.g., skin colored backgrounds, moving camera), and a larger set of individuals and gestures performed. For further evaluation, using 3D ground truth data, e.g., (Sigal and Black, 2006b) is desirable to also capture the depth accuracy of the tracking system.

Automatic and appearance independent body model acquisition provides new cues which can be of great use not only to improve the face detection, but also for person tracking or identification. The latter argument is not new, e.g., humans do not make use of inner facial features only for recognition (Sinha and Poggio, 1996). In this sense, multimodal recognition has recently become a topic of interest. Many of these approaches, however, mainly focus on face and voice modalities, neglecting kinematic properties of the body or simple features such as shirt color that do not change during an interaction session. Additional cues can provide greater robustness in the context of real world recognition applications where a robot is moving, environmental conditions are not controllable or people are not permanently facing the robot.

ACKNOWLEDGEMENTS

Work partially funded by the Spanish Ministry of Education and Science and FEDER funds (TIN2004-07087) and the Department of Universities and Research of the Canary Islands Government.

REFERENCES

Bissacco, A., Yang, M.-H., and Soatto, S. (2007). Fast human pose estimation using appearance and motion via multi-dimensional boosting regression. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*.

Brox, T., Rosenhahn, B., Kersting, U., and Cremers, D. (2006). Nonparametric density estimation for human pose tracking. In Franke, K., Mueller, R., Nickolay, B., and Schaefer, R., editors, *Pattern Recognition 2006, DAGM*, volume 4174, pages 546–555, Berlin. LNCS, Springer-Verlag, Berlin Heidelberg.

Castrillón Santana, M., Déniz Suárez, O., Hernández Tejera, M., and Guerra Artal, C. (2007). ENCARA2: Real-time detection of multiple faces at different resolutions in video streams. *Journal of*

Visual Communication and Image Representation, pages 130–140.

Fritsch, J., Lang, S., Kleinehagenbrock, M., Fink, G. A., and Sagerer, G. (2002). Improving adaptive skin color segmentation by incorporating results from face detection. In *Int. Workshop on Robot and Human Interactive Communication (ROMAN)*, pages 337–343.

Fritsch, J. and Wrede, S. (2007). *Software Engineering for Experimental Robotics*, volume 30 of *Springer Tracts in Advanced Robotics*, chapter An Integration Framework for Developing Interactive Robots, pages 291–305. Springer, Berlin.

Gavrila, D. M. (1999). The visual analysis of human movement: A survey. *Computer Vision and Image Understanding: CVIU*, 73(1):82–98.

Haasch, A., Hofemann, N., Fritsch, J., and Sagerer, G. (2005). A multi-modal object attention system for a mobile robot. In *Int. Conf. on Intelligent Robots and Systems*, pages 1499–1504.

Humanoid Animation Working Group (2007). Information technology – Computer graphics and image processing – Humanoid animation (H-Anim). http://www.h-anim.org/.

Intel (2006). Intel Open Source Computer Vision Library, v1.0. www.intel.com/research/mrl/research/opencv.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 14:201–211.

Kölsch, M. and Turk, M. (2004). Robust hand detection. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition*.

Lee, M. and Cohen, I. (2004). Human upper body pose estimation in static images. In *Proc. of European Conference on Computer Vision ECCV)*, pages 126–138.

Lömker, F., Wrede, S., Hanheide, M., and Fritsch, J. (2006). Building modular vision systems with a graphical plugin environment. In *Proc. of International Conference on Vision Systems*, page 2, St. Johns University, Manhattan, New York City, USA. IEEE.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press.

Moeslund, T. B. and Granum, E. (2001). A survey of computer vision-based human motion capture. *Computer Vision and Image Understanding: CVIU*, 81(3):231–268.

Ramanan, D. and Forsyth, D. A. (2003). Finding and tracking people from the bottom up. In *Conf. on Computer Vision and Pattern Recognition*, volume 2, pages 467–474.

S. Knoop, S. Vacek, R. D. (2006). Sensor fusion for 3d human body tracking with an articulated 3d body model. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 1686–1691, Walt Disney Resort, Orlando, Florida.

Schmidt, J., Kwolek, B., and Fritsch, J. (2006). Kernel Particle Filter for Real-Time 3D Body Tracking in Monocular Color Images. In *Proc. of Automatic Face and Gesture Recognition*, pages 567–572, Southampton, UK. IEEE.

Schneiderman, H. and Kanade, T. (2000). A statistical method for 3d object detection applied to faces and cars. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1746–1759.

Sidenbladh, H., Black, M., and Fleet, D. (2000). Stochastic tracking of 3D human figures using 2D image motion. In *Europ. Conf. on Computer Vision*, pages 702–718.

Sigal, L., Bhatia, S., Roth, S., Black, M. J., and Isard, M. (2004). Tracking loose-limbed people. In *Conf. on Computer Vision and Pattern Recognition*, volume 1, pages 421–428.

Sigal, L. and Black, M. J. (2006a). Predicting 3d people from 2d pictures. In *IV Conference on Articulated Motion and Deformable Objects - AMDO 2006*, volume 4069, pages 185–195, Mallorca, Spain. IEEE Computer Society, LNCS.

Sigal, L. and Black, M. J. (2006b). Synchronized video and motion capture dataset for evaluation of articulated human motion. Technical Report Techniacl Report CS-06-08, Brown University.

Sinha, P. and Poggio, T. (1996). I think I know that face ... *Nature*, 384(6608):384–404.

Sminchisescu, C. and Triggs, B. (2005). Mapping minima and transitions of visual models. *Int. J. of Computer Vision*, 61(1).

Stenger, B., Thayananthan, A., Torr, P., and Cipolla:, R. (2004). Hand pose estimation using hierarchical detection. In *ECCV Workshop on HCI*, pages 102–112.

Storring, M., Moeslund, T., Y.Liu, and Granum, E. (2004). Computer vision-based gesture recognition for an augmented reality interface. In *4th IASTED International Conference on VISUALIZATION, IMAGING, AND IMAGE PROCESSING*, pages 766–771.

Swain, M. J. and Ballard, D. H. (1991). Color indexing. *International Journal on Computer Vision*, 7(1):11–32.

Taycher, L., Shakhnarovich, G., Demirdjian, D., and Darrell, T. (2006). Conditional random people: Tracking humans with crfs and grid filters. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 222–229.

Urtasun, R., Fleet, D., and Fua, P. (2005). Monocular 3d tracking of the golf swing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, page 1199, San Diego.

Viola, P. and Jones, M. J. (2004). Robust real-time face detection. *International Journal of Computer Vision*, 57(2):151–173.