Hardware-assisted Multiple Object Tracking for Human-Robot-Interaction

Claus Lenz, Giorgio Panin, Thorsten Röder, Martin Wojtczyk,
and Alois Knoll
Robotics and Embedded Systems Lab
Boltzmannstrasse 3
85748 Garching bei München
Technische Universität München
{lenz,panin,roeder,wojtczyk,knoll}@in.tum.de

Categories and Subject Descriptors

I.2.10 [**Vision and Scene Understanding**]: Motion; I.4 [**IMAGE PROCESSING AND COMPUTER VISION**]: Applications

General Terms

Algorithms

Keywords

model-based tracking, GPU, HRI, joint-action

1. EXTENDED ABSTRACT

At the moment, the collaboration of human and robot is mainly based on a master-slave level with a human teleoperating the robot or programming it off-line allowing the robot to execute only static tasks. In industrial production the limitations make a collaboration at the moment nearly impossible, because of the needed safety for the human worker. Therefore, to ensure safety, the workspaces of humans and robots are strictly separated in time or in space. This workspace splitting does not take advantage of the potential for humans and robots to work together as a team, where each member has the possibility to actively assume control and contribute towards solving a given task based on their capabilities. Such a mixed-initiative system supports a spectrum of control levels, allowing the human and robot to support each other in different ways, as needs and capabilities change throughout a task [4]. With the subsequent flexibility and adaptability of a human-robot collaboration team, production scenarios in permanently changing environments as well as the manufacturing of highly customized products become possible.

One step towards the goal of an efficient and safe collaboration between human and robot is to give the robot "eyes" to detect and track the human worker in order to avoid collisions, to figure out what the human is doing, and to be able to hand over objects. In this paper, we propose a hardware-assisted multiple-object tracking system for human-robot-interaction based on particle filters and pixel-level likelihoods. The proposed method computes for each multi-target particle a full hypothesis-map through the rendering engine of the graphics card, and compares it with the underlying binary map of the image-preprocessing on the fragment shader of the GPU. The approach is formulated in a generic way with respect to the segmentation method, the object shape, and the number of targets to cover a magnitude of tasks within human-robot interaction. It is a further development of our work presented in [3] and will be used on our demonstration platform named *JAHIR - Joint-Action for Humans and Industrial Robots* [2] to enable such a fruitful and safe collaboration of human and industrial robot.

HRI'09, March 11–13, 2009, La Jolla, California, USA.
ACM 978-1-60558-404-1/09/03.

2. THE MULTIPLE-OBJECT TRACKING SYSTEM

To track multiple objects at the same time with only one filter the standard sampling-importance-resampling (SIR) scheme [1] was extended to particles with multiple object hypothesis (MOSIR). Therefore, one particle contains a set of $i = 1 \dots I$ targets and forms a complete hypothesis scene.

$$\left\{ \{s^n_{t,i}\}^I_{i=1}, \pi^n_t \right\};\ n = 1 \dots N \tag{1}$$

The weight π_n of each particle $n = 1 \dots N$ is computed by comparing its hypothesis scene $\{s^n_{t,i}\}^I_{i=1}$ with the current measurement z_t, obtained by a Gaussian-Mixture-Modell-based (GMM) segmentation in the pre-processing step resulting in a binary image (see Fig. 2).

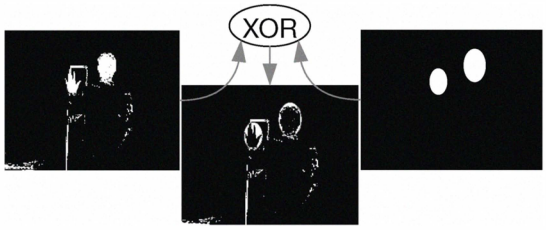

Figure 2: Left: Segmented image. Right: generated multi-target shadow hypothesis. Middle: pixel-level residual.

For pixel-level matching, each target object needs to be projected to the camera plane and a *shadow-map* h^n_t of the whole (hypothesis) scene is to be generated. The shadow-map provides a binary image representing the expected measurement for an ideal, noise-free segmentation under a given scene $\{s^n_{t,i}\}^I_{i=1}$ (Fig. 2, right). The expected scene is then

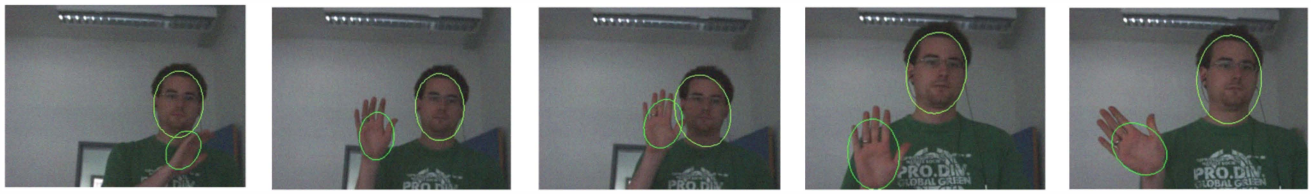

Figure 1: Simultaneous tracking of hand and face with an elliptic model.

compared to the real measurement (Fig. 2, left) by a sum of square differences (SSD) cost function, that is in binary images h_t^n, z_t equivalent to a pixel-wise XOR (Fig. 2, middle) followed by a sum of the non-zero pixels.

The computation of h^n is very expensive if performed on the CPU, and moreover only limited pixel parallelization can be exploited while comparing it with the measurement z_t. Therefore, we have implemented these operations on the GPU, using at the same time the power of the rendering engine, and the parallel pixel-pipelines.

The residual values are normalized and the likelihood is evaluated with a Gaussian likelihood model:

$$\pi_n = P(z_t|\{s_{t,i}^n\}_{i=1}^I) = k \cdot \exp(-\frac{e_t^n}{2r^2}) \qquad (2)$$

with a suitable measurement variance r, providing the new particle weights π_n, afterwards normalized so that $\sum_n \pi_n = 1$. Deterministic resampling of the particle set [1] is applied after each update, in order to keep a well-distributed particle set.

For all details concerning the GPU-assisted segmentation (Fig. 2, left), the generation of the hypothesis-map and the matching on the GPU, we kindly refer to [3].

3. EXPERIMENTAL RESULTS

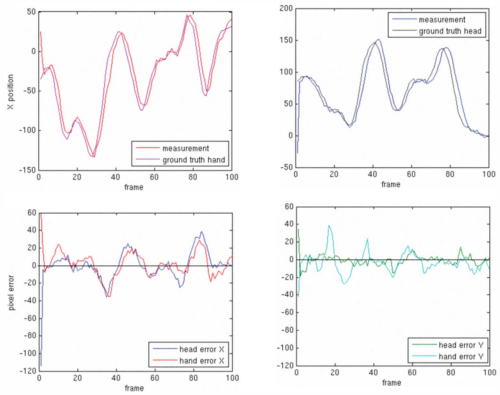

Figure 3: upper line: Comparison between the estimated trajectory in x direction for the two objects. lower line: resulting error between measured track and corresponding ground truth.

As already emphasized in the previous descriptions, our methodology can be applied to different visual modalities, with different object shapes, degrees of freedom, and numbers of targets. The results presented in this work deal with the simultaneous tracking of multiple skin colored objects, because in most cases of a human-robot collaboration, these body parts are of most interest. The object shape was defined as an ellipsoid approximating the shape of a human hand and head as well.

For the experiments, an Intel dual-core machine with programmable graphics card (NVidia GeForce 8800) was used, with input images captured from a standard FireWire camera.

The GMM model was built from a training data set of labeled skin-pixels, and consists of $K = 2$ mixture components, as in [5]. The processing speed for image segmentation, including the conversion from RGB to HSV, has been 20 times faster on the GPU, already providing a big advantage for tracking.

The upper diagrams of Figure 3 show the estimated trajectory of the hand (left) and the head (right) in x direction. After the uniform initialization the filter condenses to the wanted objects and tracks them throughout the sequence of 100 frames. Although we used a very basic motion model, the trajectories are very equal. This can also be seen in the lower diagram of Figure 3, which shows the total pixel error for the two objects in x and y direction compared to hand-annotated hand and head position for each frame.

Output results of the tracking can be seen in Figure 1 illustrating the tracking procedure with selected screenshots of the output. The tracking was initialized with an uniformly distributed particle set around the image center.

4. ACKNOWLEDGEMENTS

This work is partly supported by the German Research Council (DFG), under the excellence initiative cluster *CoTeSys - Cognition for Technical Systems* (http://www.cotesys.org).

5. REFERENCES

[1] M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *International Journal of Computer Vision (IJCV)*, 29(1):5–28, 1998.

[2] C. Lenz, S. Nair, A. Knoll, W. Rösel, J. Gast, F. Wallhoff, and M. Rickert. Joint-action for humans and industrial robots for assembly tasks. In *RO-MAN 08: Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication*, München, Germany, 2008. IEEE Robotic and Automation Society.

[3] C. Lenz, G. Panin, and A. Knoll. A GPU-accelerated particle filter with pixel-level likelihood. In *International Workshop on Vision, Modeling and Visualization (VMV)*, Konstanz, Germany, Oct. 2008.

[4] J. L. Marble, D. J. Bruemmer, D. A. Few, and D. D. Dudenhoeffer. Evaluation of supervisory vs. peer-peer interaction with human-robot teams. In *HICSS '04: Proceedings of the Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04) - Track 5*, page 50130.2, Washington, DC, USA, 2004. IEEE Computer Society.

[5] M. Yang and N. Ahuja. Gaussian Mixture Model for Human Skin Color and Its Application in Image and Video Databases. In *Proc. SPIE: Storage and Retrieval for Image and Video Databases VII*, volume 3656, pages 458–466, 1999.