# Tracking Articulated Structures in Stereo Image Sequences

Nebojsa Jojic, Thomas S. Huang
University of Illinois
Beckman Institute and ECE Department
405 N. Mathews Ave, Urbana, IL 61801
Email: jojic@ifp.uiuc.edu

Matthew Turk
Microsoft Research
One Microsoft Way Redmond, WA

*Abstract* — **In this paper, we present an algorithm for real time 3-D tracking of articulated structures in stereo image sequences. These sequences can be captured by an inexpensive commercially available system that also computes the dense disparity map in real time. In our algorithm, the tracked object is modeled as a set of articulated 3D blobs, each adhering to a Gaussian distribution. Classification of the disparity map pixels into the segments of the articulated object is based on the maximum likelihood principle with an additional mechanism for filling the missing data created by self-occlusions. The articulation constraints are enforced through an Extended Kalman Filter, which can also be used to model the dynamics of the tracked object.**

## I. Introduction

In the past, there were several approaches to human body tracking, mainly based on single image sequences, and therefore often limited to tracking the 2-D motion only. To satisfy the real-time requirement, simplified statistical human body models have been proposed (see [1], for example). In [2], the output of two 2-D trackers was combined to achieve full 3-D tracking. However, only the hands and the head of the user were tracked. Knowledge of the dynamics properties of the human motion is believed to be helpful for tracking [2].

In this paper we present an algorithm that uses the dense disparity map computed from the input from two cameras to estimate the posture of an articulated structure at each time frame.

## II. Articulated statistical model

The articulated model consists of several links, which are linked at the joints. Each link is assumed to produce a Gaussian distribution of 3-D points corresponding to the frontal parts of the tracked object. In the rest of the paper, we refer to these statistical models of individual parts as blobs.

The full model can be described by the orientations of all links and the center of one of them, providing that the joint positions in the local, link-based coordinate systems are known, as well as the size parameters of the blobs. From the model parameters, we compute for each link the mean and the covariance matrix of the corresponding 3-D blob. Providing that the articulated model parameters accurately capture the posture of the imaged object, the blob statistical model can then be used to label the pixels in the dense disparity map, using the maximum likelihood classifier. In this way we segment the images into different body parts (for example, head, torso, lower and upper arms...)

On the other hand, given the labeled disparity map, the means and covariance matrices of all blobs can be re-estimated. An Extended Kalman Filter (EKF) uses these estimated values as measurements which are non-linearly related to the orientations of the links and the reference link center. In these measurement equations, the size parameters of the blobs (eigen values of the covariance matrices) are assumed to remain constant and the means and orientations of the covariance matrices of the distributions are assumed to satisfy the kinematic chain rules. Thus, EKF prevents the blobs from diminishing in size and enforces the link constraints.

## III. Occlusion detection and handling

When one of the body parts is occluding the other in the captured image sequence, the blob model of the (partially) occluded part will have incomplete data for estimation of the mean and the covariance of the Gaussian model.

To prevent this, it is necessary to detect such situations. After the disparity map has been labeled, we know for each pixel which blob it belongs to. For all other blobs, we can use the statistical model to estimate the most likely depth given the image coordinates. The occlusion is detected when the substitution of the measured depth by the estimated optimal depth for a potentially occluded blob increases its likelihood so that it becomes close to the likelihood of the winner blob. Another condition is that the optimal depth is larger than the measured depth.

Once the occlusion is detected, the missing data can be filled with optimal depth values, which in combination with the constraints in the EKF handles occlusion sufficiently well.

## IV. Experiments

In our experiments we used 160x120 disparity maps obtained from two cameras 8cm apart in real time. We tested our algorithm on tracking the upper human body using a model consisting of head, torso, upper arm and lower arm blobs, which were linked appropriately. The tracker handled occlusions well and ran at about 5-10 frames/second on a 333MHz Pentium II PC.

## References

[1] C. R. Wren, A. Azarbayejani, T. Darrell, A. P. Pentland, "Pfinder: real-time tracking of the human body," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.19, no.7, July 1997, pp.780-5.

[2] C. R. Wren, A. Pentland, "Dynamic models of human motion," *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition* 1998, pp.22-7.