

# A BAYESIAN FORMULATION FOR 3D ARTICULATED UPPER BODY SEGMENTATION AND TRACKING FROM DENSE DISPARITY MAPS

Robert D. Cavin<sup>1</sup>, Ara V. Nefian<sup>1</sup> and Navin Goel<sup>2</sup>

<sup>1</sup>Microprocessor Research Labs, Intel Corporation  
 {robert.d.cavin, ara.nefian}@intel.com

<sup>2</sup>Department of Computer Science, University of Nevada at Reno  
 goel@cs.unr.edu

## ABSTRACT

This paper describes a Bayesian network for 3D articulated upper body segmentation and tracking from video sequences for which both color and depth information are available. In our upper body model the joints are represented as the parent nodes of the body components nodes which include the head, torso or arms. The upper body components are modeled using a set of planar, linear and Gaussian density functions. The model described in this paper segments and tracks accurately the upper body in different illumination conditions and in the presence of partial occlusions and self occlusions. In addition the current approach allows for automatic segmentation of the upper body without any human intervention allowing for further use of the system in hand gesture or human activity recognition.

## 1. INTRODUCTION

The segmentation and tracking of the human body captivates the attention of the research community in computer vision due to their applications in automatic human activity understanding, gesture recognition, robotics or industrial control. Tracking systems that use video sequences captured with one camera make use of the color information [9], complex geometry of the body [1] or the dynamics of human motion [2]. Such systems often require a constraint background, a priori knowledge of the scale information, and human intervention in the initialization process. Human body tracking systems from multiple cameras [8, 6] have demonstrated their improved performances in dealing with variations in scale and body occlusions. However, by their nature these methods are restricted to environments where a complex system with multiple cameras is available. Recently, consumer-level stereo cameras are becoming more commonplace and the performance of personal computers is approaching the threshold where stereo computation can be done at reasonable frame rates.<sup>1</sup> The use of dense disparity maps together with colors increases considerably the robustness of our system to variations in illumination conditions. It also reduces the inherent depth ambiguity present in 2D images and therefore enables accurate segmentation under partial occlusions and self-occlusions. As a result, robust techniques based on the use of stereo images and the depth information, have been already considered for various applications such as pointing [5], tracking

<sup>1</sup>For example, stereo camera produced and sold by Point Grey Research, Inc. running on 1.5GHz Pentium<sup>TM</sup> 4 can compute 320x240 disparity maps at 11 frames per second.

[4], or static gesture recognition [3]. In this paper we introduce a Bayesian model for the upper body from dense disparity maps (Section 2), describe the parameter estimation of the model used in segmentation and tracking (Section 3) and present the experimental results (Section 4) on an extensive hand gesture database.

## 2. THE UPPER BODY MODEL

The first step towards the accurate segmentation and tracking of the upper body is the detection of the foreground pixels. The foreground pixels are those for which 3D position is available and which are closer to the camera than a fixed distance threshold. Experimental results [5, 4, 7] showed that the disparity-based segmentation is significantly more robust to non-stationary background and variations in illumination than the color-based segmentation methods. The statistical upper body model for the foreground pixels is illustrated as a Bayesian network in Figure 1. The nodes

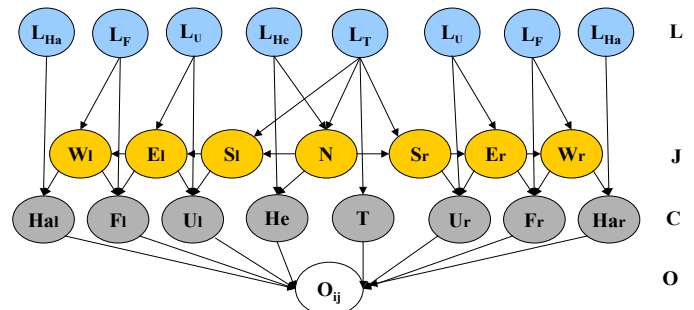


Fig. 1. The upper body model.

of the network represent the body components **C**, the joints **J**, the anthropological measures **L** and the observation nodes  $O_{ij}$ . The body components consist of the torso  $T$ , head  $He$ , left and right upper arms  $U_l$  and  $U_r$ , left and right forearms  $F_l$  and  $F_r$  and the left and right hands  $Ha_l$  and  $Ha_r$ . The **J** nodes describe the 3D positions of the upper body joints such as the neck  $N$ , left and right shoulders,  $S_r$  and  $S_l$ , the left and right elbows  $E_l$  and  $E_r$  and the left and right wrists  $W_l$  and  $W_r$ . The anthropological measure nodes **L** represent the 3D size of each body component such as the torso  $L_T$ , head  $L_{He}$ , left and right upper arms,  $L_{U_l}$  and  $L_{U_r}$ , left and right forearms,  $L_{F_l}$  and  $L_{F_r}$ , and left and right hands,  $L_{Ha_l}$  and  $L_{Ha_r}$ . The foreground observation vector  $O_{ij}$  corresponding to pixel in row  $i$  and column  $j$  consists of the three

dimensional position of the pixel as obtained from the disparity maps  $\mathbf{O}_{ij}^d = [x_{ij}, y_{ij}, z_{ij}]^T$  and its color in the image space  $\mathbf{O}_{ij}^c$ . We found that using the hue value extracted from the HSV color space is sufficient to robustly describe the data and keeps a low computational complexity of the model. The probability of the observation vectors  $\mathbf{O}_{ij}$  given the upper body model  $\Omega$  is obtained as

$$P(\mathbf{O}_{ij}|\Omega) = \sum_{\mathbf{J}, \mathbf{C}, \mathbf{L}, q_{ij}} P(\mathbf{O}_{ij}, q_{ij}, \mathbf{J}, \mathbf{C}, \mathbf{L}|\Omega) \quad (1)$$

where  $q_{ij}$  is the value or state of the  $\mathbf{C}$  node corresponding to  $\mathbf{O}_{ij}$  and

$$P(\mathbf{O}_{ij}, q_{ij}, \mathbf{J}, \mathbf{C}, \mathbf{L}|\Omega) = P(\mathbf{O}_{ij}|q_{ij}, \mathbf{C}, \mathbf{L})P(q_{ij}|\mathbf{J}, \mathbf{C}, \mathbf{L})P(\mathbf{J}, \mathbf{C}|\mathbf{L})P(\mathbf{L}) \quad (2)$$

Each of the factors of the above equation will be explained in more detail. First, it is natural to assume that the values of the nodes  $\mathbf{L}$  do not change over time for the same person. In addition, in this paper we assume that these values are also known a priori. This simplification significantly increases the speed of the segmentation by allowing  $P(\mathbf{L}) = \text{const}$ .

Since the color distribution and the 3D position can be considered independent random variables, the probability of the observation vectors  $\mathbf{O}_{ij}$  can be decomposed as:

$$P(\mathbf{O}_{ij}|q_{ij}, \mathbf{L}) = P(\mathbf{O}_{ij}^d|q_{ij}, \mathbf{L})P(\mathbf{O}_{ij}^c|q_{ij}) \quad (3)$$

In our method,  $P(\mathbf{O}_{ij}^c|q_{ij})$  is described by a uniform distribution over the entire range of hue values [1, ..., 255] for the torso and arms and is a Gaussian density function for the head and hand components. The nodes corresponding to the components of the upper body model are defined by a set of Gaussian density functions for the head and hands, *linear* density functions for the upper arms and forearms, and a planar density function [7] for the torso. The probability of the observation vector  $\mathbf{O}_{ij}^d$  given the head component  $H_e = \{\mu_{H_e}, \mathbf{C}_{H_e}\}$  is:

$$P(\mathbf{O}_{ij}^d|H_e, L_{H_e}) = K_{H_e} \mathbf{N}(\mathbf{O}_{ij}|\underline{\mu}_{H_e}, \mathbf{C}_{H_e}) \times \mathbf{U}(\mathbf{O}_{ij}|\underline{\mu}_{H_e}, L_{H_e}) \quad (4)$$

where  $\mathbf{N}$  is a Gaussian density function with mean  $\underline{\mu}_{H_e}$  and covariance  $\mathbf{C}_{H_e}$ ,  $\mathbf{U}$  is a tridimensional uniform distribution with mean  $\underline{\mu}_{H_e}$ , and support  $L_{H_e}$ , and  $K_{H_e}$  is a normalization constant. A similar pdf is obtained for the left and right hand components ( $H_{a_l}$  and  $H_{a_r}$ ).

The probability of the observation  $\mathbf{O}_{ij}^d$  given the torso component  $T = \{a, b, c, \sigma_z\}$  is defined as:

$$P(\mathbf{O}_{ij}^d|T, L_T) = K_T \frac{1}{\sqrt{2\pi\sigma_z^2}} \exp\left(-\frac{(z_{ij} - \pi_{ij})^2}{2\sigma_z^2}\right) \times \mathbf{U}([x_{ij}, y_{ij}]^T | [\mu_{T_x}, \mu_{T_y}]^T, [L_{T_x}, L_{T_y}]^T) \quad (5)$$

where  $\pi_{ij} = ax_{ij} + by_{ij} + c$ , is the mean plane and  $\mu_T$  is the mean of torso pdf,  $L_T$  is the 3D anthropological measure associated with the torso and  $K_T$  is a normalization constant. Note that the planar pdf measures the distance on the  $z$  axis, rather than orthogonal distance between the observation and the mean plane. However, under the common assumption that the normal to the user's body is facing the camera, the  $z$  distance is a good approximation of the orthogonal distance.

The left and right upper arms and forearms are described by a set of *linear* probability density functions. A linear pdf can be seen as normal density function for which the mean is represented by a line segment between two points. Without loss of generality, let the spherical coordinates of the points be the origin and  $(r_{max}, \theta, \phi)$  respectively. More formally a linear pdf with the parameters  $A = \{r_{max}, \theta, \phi, \sigma\}$  is defined as:

$$P(\mathbf{O}_{ij}^d|A) = \frac{K_A}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(\mathbf{O}_{ij}^d - \lambda_{ij})^2}{2\sigma^2}\right) \times \mathbf{U}\left(\mathbf{O}_{ij}^d \left| \frac{r_{max}}{2}, \frac{r_{max}}{2}\right.\right) \quad (6)$$

where  $\lambda_{ij} = (r, \theta, \phi)$  is the projection of the observation  $\mathbf{O}_{ij}$  on the line segment between the two points and  $K_A$  is a normalization constant. Note that the linear pdf uses the orthogonal distance between an observation and the mean line segment, which allows for arbitrary arm movement relative to the camera.

Let  $J_c$  and  $J_p$  be two adjacent joints (e.g. wrist and elbow) that represent the parent nodes of a body component described by a linear pdf (e.g. forearm). Then, the conditional probability of the linear pdf parameters is described by

$$P(A|J_c, J_p) = \delta([r_{max}, \phi, \theta]^T - [r_{J_c}, \phi_{J_c}, \theta_{J_c}]^T) \quad (7)$$

where  $(r_{J_c}, \theta_{J_c}, \phi_{J_c})$  are the spherical coordinates of  $J_c$  with the origin in  $J_p$ .

The conditional probability of a joint  $J_c$  given its parent joint  $J_p$  and the anthropological measure  $L$  is given by:

$$P(J_c|J_p, L) = \begin{cases} K_{J_c} & \text{if } \theta_{J_c} \in [\theta_{min}, \theta_{max}] \\ & \text{and } \phi_{J_c} \in [\phi_{min}, \phi_{max}] \\ & \text{and } r_{J_c} = L \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

where  $K_{J_c}$  is a normalization constant,  $\phi_{min}$ ,  $\theta_{min}$  and  $\phi_{max}$ ,  $\theta_{max}$  represent the minimum and maximum values of parameters  $\phi_{J_c}$  and  $\theta_{J_c}$  for natural human motion respectively. Equation 8 also describes the conditional probability of the head and hands parameters  $\mu_{H_e}$ ,  $\mu_{H_{a_l}}$  and  $\mu_{H_{a_r}}$  given their parent joints i.e. the neck  $N$ , left and right wrist,  $W_l$  and  $W_r$  respectively.

### 3. THE UPPER BODY SEGMENTATION AND TRACKING

For the set of independent foreground observations,

$$P(\mathbf{O}_F|\Omega) = \prod_{all\ ij \in F} P(\mathbf{O}_{ij}|\Omega) \quad (9)$$

where  $F$  is the set of foreground pixels and  $\mathbf{O}_F = \{\mathbf{O}_{ij} | ij \in F\}$  represents the sequence of foreground observation vectors. With the approximation to Equation 1

$$P(\mathbf{O}_{ij}|\Omega) \approx \max_{\mathbf{J}, \mathbf{L}, q_{ij}} P(\mathbf{O}_{ij}, q_{ij}, \mathbf{J}, \mathbf{C}, \mathbf{L}|\Omega) \quad (10)$$

and from Equation 9,

$$P(\mathbf{O}_F|\Omega) \approx \max_{\mathbf{J}, \mathbf{C}, \mathbf{L}, \mathbf{Q}_F} P(\mathbf{O}_F, \mathbf{Q}_F, \mathbf{J}, \mathbf{C}, \mathbf{L}) \quad (11)$$

where  $\mathbf{Q}_F = \{q_{ij} | j \in F\}$ . Equation 11 allows for efficient computation in logarithmic representation of the observation likelihoods defined in Section 2. Therefore, for a fixed set of anthropological measures, the initialization problem reduces to finding

$$\{\mathbf{J}, \mathbf{C}, \mathbf{Q}_F\} = \arg \max_{\mathbf{Q}_F, \mathbf{J}, \mathbf{C}} P(\mathbf{O}_F, \mathbf{J}, \mathbf{C}, \mathbf{Q}_F) \quad (12)$$

However, finding the best set of joints  $\mathbf{J}$ , components  $\mathbf{C}$  and state sequence  $\mathbf{Q}_F$  that maximizes  $P(\mathbf{O}_F, \mathbf{J}, \mathbf{C}, \mathbf{Q}_F)$  remains a complex problem. In this paper, rather than computing the best overall set of joints and components, we separate the segmentation problem into two stages and reduce its complexity. First we segment the body components  $B = \{He, T\}$  and the body joints  $\mathbf{J}_B = \{N, S_l, S_r\}$  and assign all the remaining pixels the background. In the second stage, the pixels not assigned to the body components  $B$  are used to determine the parameters of the arm components  $A = \{U_l, U_r, F_l, F_r, Ha_l, Ha_r\}$  and the arm joints  $\mathbf{J}_A = \{E_l, E_r, W_l, W_r\}$ . The stages of our segmentation algorithm are described below in more detail.

**Stage 1.** In this stage of the algorithm we assume that there is only one visible user in the image and that the torso is largest visible body component. Assuming for simplicity that the head is in a vertical position the initialization starts from an arbitrary position of the neck. During tracking, the initial neck position is estimated from the position of the neck found in the previous frame. Within each frame the parameters of the components  $B$  and the position of the corresponding joints  $\mathbf{J}_B$  are obtained through the following iteration:

1. The parameters of the torso are estimated from all observation vectors below the horizontal plane that contains the neck using the EM approach presented in [7].
2. Assuming the head in vertical position i.e.  $\mu_{He_x} = \mu_{Tx} = N_x$  and  $\mu_{He_y} = N_y + L_{He_y}/2$ , the head parameters are estimated from the observation vectors in a 3D bounding box centered on  $\mu_{He} = [\mu_{Tx}, N_y + L_{He_y}/2, a\mu_{Tx} + b(N_y + L_{He_y}/2) + c]^T$  using EM [7].

3. Compute

$$\tilde{q}_{ij} = \arg \max_{q_{ij} \in B} \log P(\mathbf{O}_{ij}, q_{ij}, \mathbf{J}_B, \mathbf{C})$$

and

$$\log \tilde{P}(\mathbf{O}_F, \mathbf{Q}_F, \mathbf{J}_B, \mathbf{C}) = \sum_{ij \in F} \log P(\mathbf{O}_{ij}, q_{ij}, \mathbf{J}_B, \mathbf{C}).$$

4. Re estimate the position of the joint parameters  $\mathbf{J}_B$  such that

$$\begin{aligned} N &= [\mu_{He_x}, \mu_{He_y} - \frac{L_{He_y}}{2}, aN_x + bN_y + c]^T \\ S_l &= [\mu_{Tx} + \frac{L_{Tx}}{2}, N_y, aS_{lx} + bS_{ly} + c]^T \\ S_r &= [\mu_{Tx} - \frac{L_{Tx}}{2}, N_y, aS_{rx} + bS_{ry} + c]^T \end{aligned}$$

5. Repeat steps 1-4 for the new set of joints until  $\log P(\mathbf{O}_F, \mathbf{Q}_F, \mathbf{J}_B, \mathbf{C})$  at consecutive iteration falls under the convergence threshold.

**Stage 2.** In this stage we select the best position of the arms from a set of possible arm configurations. An arm configuration consists of all the joints of the arms and the mean of the hands distributions. The selection criteria is the maximization of the arms likelihood over all the pixels not assigned in the previous stage to the head or torso. The position of a joint  $J_c$  or hand center in a valid arm configuration is obtained from the significant values of the conditional density shown in Equation 8, i.e.

$$J_c = J_p + [L \cos \theta_{J_c}, L \sin \theta_{J_c} \cos \phi_{J_c}, L \sin \theta_{J_c} \sin \phi_{J_c}]^T$$

where  $L$  is a constant describing the anthropological measure of the arm component and  $J_p$  is the parent joint. For a set of valid arm configurations the values  $\phi_{J_c}, \theta_{J_c}$  are chosen equally spaced in the  $[(\phi_{min}, \theta_{min})^T, (\phi_{max}, \theta_{max})^T]$  range. For the initialization of the model we selected 16, 18 and five  $\phi_{J_c}, \theta_{J_c}$  values for the elbows, wrists and center of the hands respectively. In tracking at time  $t$ , the position of a joint in a valid configuration,  $J_c(t)$  is determined from five  $\phi_{J_c}, \theta_{J_c}$  samples equally spaced around the  $\phi_{J_c}(t-1), \theta_{J_c}(t-1)$  values estimated from the previous frame and according to:

$$\begin{aligned} J_c(t) &= J_p(t) + [L \cos \theta_{J_c}(t-1), L \sin \theta_{J_c}(t-1) \\ &\quad \cos \phi_{J_c}(t-1), L \sin \theta_{J_c}(t-1) \sin \phi_{J_c}(t-1)]^T \end{aligned}$$

The arm joint segmentation is described by the following steps:

1. For each possible arm joint configuration we estimate the mean of the linear density functions corresponding to the upper arms and forearms, and the mean of the normal pdf for the hands.
2. For each joint configuration we determine the best state assignment of the observation vectors

$$\tilde{q}_{ij}(\mathbf{J}_A) = \arg \max_{q_{ij} \in A} \log P(\mathbf{O}_{ij}, \mathbf{J}_A, \mathbf{C}, q_{ij})$$

$$\tilde{\mathbf{Q}}_A(\mathbf{J}_A) = \{\tilde{q}_{ij}(\mathbf{J}_A) | ij \in A\}$$

and the observation log likelihood

$$\log \tilde{P}(\mathbf{O}_A, \mathbf{Q}_A, \mathbf{J}_A, \mathbf{C}) = \sum_{ij \in A} \log P(\mathbf{O}_{ij}, \tilde{q}_{ij}(\mathbf{J}_A), \mathbf{J}_A, \mathbf{C})$$

where  $\mathbf{O}_A$ , and  $\mathbf{Q}_A$  are the observation vectors and the states corresponding the foreground pixels not assigned to the head or torso in the first stage of the algorithm.

3. Find the highest likelihood over all joint configurations  $\mathbf{J}_A$

$$\log P^*(\mathbf{O}_A, \mathbf{Q}_A, \mathbf{J}_A, \mathbf{C}) = \max_{\mathbf{J}_A} \log \tilde{P}(\mathbf{O}_A, \mathbf{Q}_A, \mathbf{J}_A, \mathbf{C})$$

and determine the best set of joints and the corresponding best state assignment.

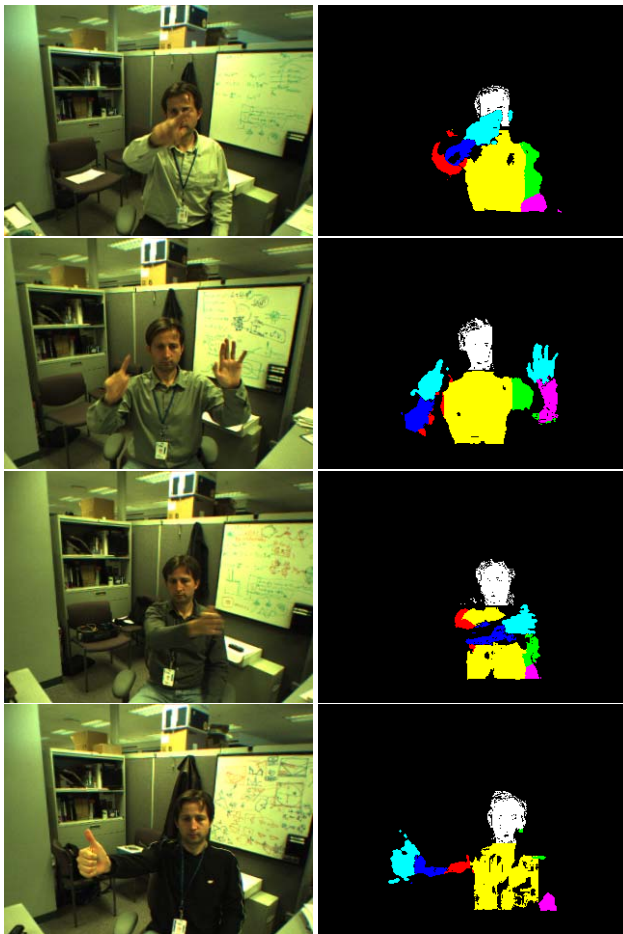
$$\mathbf{J}_A^* = \arg \max_{\mathbf{J}_A} \log P^*(\mathbf{O}_A, \mathbf{Q}_A, \mathbf{J}_A, \mathbf{C})$$

$$\mathbf{Q}_A^* = \tilde{\mathbf{Q}}_A(\mathbf{J}_A^*)$$

## 4. EXPERIMENTAL RESULTS

A DigiClops<sup>TM</sup> camera system [10] was used to acquire stereo sequences of 16 static and dynamic gestures [7]. The dynamic gestures represent translations and rotations in the image plane and in

the plane perpendicular to the image plane. The sequences were captured over the period of two months and therefore vary in illumination and environment settings. All the sequences were written to the disk drive at 15 fps at the resolution of 320x240. The disparity maps for all the frames were computed off line. In total, we captured 640 sequences of dynamic and static gestures, i.e., 40 examples for each gesture. Figure 2 shows typical segmentation results and displays each of the segmented body components in different gray levels. Dark regions represent regions not assigned to the upper body components or regions for which no depth information is available. The results show that the algorithm presented in this paper can successfully segment the body components in the presence of self occlusions and various 3D orientations. Further analysis in the color space of the pixels with no depth information will increase the accuracy of the current segmentation.



**Fig. 2.** Examples of images of the upper body (left) and the corresponding segmentation results (right)

## 5. CONCLUSIONS AND FUTURE WORK

This paper presents a Bayesian network used to segment and track the articulated 3D upper body. The model uses the color and 3D information generated by a stereo camera. Unlike other tracking systems which require a user guided initialization, our approach

for upper body segmentation makes use of a minimal set of assumptions of the relative position of the user to the camera to initialize automatically. The use of the dense disparity maps in our statistical framework improves the performance over a system using color and pixel coordinates alone by making the segmentation robust to illumination changes and by providing the full three dimensional motion data for our model of the upper body. We believe that the high accuracy of the upper body segmentation and tracking system presented in this paper will lead to interesting and stimulating work on natural unobtrusive vision-based user interfaces. Further research will be directed towards building a gesture recognition system using both upper body pose and image data that is segmented to the hands.

## 6. REFERENCES

- [1] C. Barron and I. Kakadiaris. Estimating anthropometry and pose from a single image. In *Computer Vision and Pattern Recognition*, pages 669–676, 2000.
- [2] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *Computer Vision and Pattern Recognition*, pages 8–15, 1998.
- [3] R. Grzeszczuk, G. Bradski, M.H. Chu, and J.Y. Bouguet. Stereo based gesture recognition invariant to 3D pose and lighting. In *International Conference on Computer Vision and Pattern Recognition*, pages 826–833, 2000.
- [4] N. Jovic, B. Brumitt, B. Meyers, S. Harris, and T. Huang. Tracking self-occluding articulated objects in dense disparity maps. In *International Conference on Computer Vision*, pages 123–130, 1999.
- [5] N. Jovic, B. Brumitt, B. Meyers, S. Harris, and T. Huang. Detection and estimation of pointing gestures in dense disparity maps. In *International Conference on Face and Gesture Recognition*, pages 468–475, 2000.
- [6] I. Kakadiaris and D. Metaxas. Model-based estimation of 3D human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1453–1459, 2000.
- [7] A. V. Nefian, R. Grzeszczuk, and V. Erubimov. A statistical upper body model for 3D static and dynamic gesture recognition from stereo sequences. In *IEEE International Conference on Image Processing*, pages 601–607, 2001.
- [8] H. Sidenbladh, F. De La Torre, and M. J. Black. A framework for modeling the appearance of 3D articulated figures. In *Automatic Face and Gesture Recognition*, pages 368–375, 2000.
- [9] C. Wren, A. Azerbayejani, T. Darell, and A. Pentland. Pfunder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:780–785, July 1997.
- [10] Point Grey Research, Digiclops Stereo System, <http://www.ptgrey.com/products/digiclops>.