Learning and synthesizing human body motion and posture

Rómer Rosales and Stan Sclaroff
Boston University, Computer Science Department
111 Cummington St., Boston, MA 02215
email:{rrosales,sclaroff}@bu.edu

Abstract

A novel approach is presented for estimating human body posture and motion from a video sequence. Human pose is defined as the instantaneous image plane configuration of a single articulated body in terms of the position of a predetermined set of joints. First, statistical segmentation of the human bodies from the background is performed and low-level visual features are found given the segmented body shape. The goal is to be able to map these visual features to body configurations. Given a set of body motion sequences for training, a set of clusters is built in which each has statistically similar configurations. This unsupervised task is done using the Expectation Maximization algorithm. Then, for each of the clusters, a neural network is trained to build this mapping. Clustering body configurations improves the mapping accuracy. Given new visual features, a mapping from each cluster is performed providing a set of possible poses. From this set, the most likely pose is extracted given the learned probability distribution and the visual feature similarity between hypothesis and input. Performance of the system is characterized using a new set of known body postures, showing promising results.

1 Introduction

In recent years, there has been a great deal of interest in methods for tracking and analysis of human body motion by computer [1, 3, 18, 14, 20, 9, 12, 7, 15, 17, 11, 16, 4]. Effective solutions would lead to breakthroughs in areas such as video surveillance, human motion recognition, ergonomics, motion performance measurement, human computer interfaces, virtual reality, computer animation, and robot navigation, among others.

It is clear that if the basic structure of the tracked body (its configuration) is reconstructed, motion analysis would be greatly simplified. In our everyday life, humans can easily estimate body part location and structure from relatively low-resolution images of the projected 3D world (*e.g.*,watching a video). Unfortunately, this problem is inherently difficult for a computer. Despite research attention, only in very well controlled situations, normally not useful for interesting applications, have researchers been able to obtain relatively satisfactory results. Finding the mapping between low-level image features and body configurations is highly complex and ambiguous. The difficulty stems from the number of degrees of freedom in the human body, the complex underlying probability distribution, ambiguities in the projection of human motion onto the image plane, self-occlusion, insufficient temporal or spatial resolution, etc.

In this paper, we present a novel approach for the estimation of human body pose and motion given a single 2D view of a scene containing unoccluded bodies. Human pose is defined as the instantaneous image plane configuration of a single articulated body in terms of the position of a predetermined set of joints. Given a set of body motion sequences for training, a set of clusters is built in which each has statistically similar configurations. Then, for each of the clusters, a neural network is trained to build this mapping. Experiments show that clustering body configurations improves the mapping accuracy. Given new visual features, a mapping from each cluster is performed providing a set of possible poses. From this set, the most likely pose is extracted given the learned probability distribution and the visual feature similarity between hypothesis and input.

The approach consists of learning how specific body classes (*e.g.*,human bodies) are configured. This is done by observing examples of body configurations. Because body configurations have some underlying structure, this can reduce considerably the volume of the space of possible configurations. We will show how we can efficiently learn from data sets of body configurations, and using this prior knowledge, how to map low-level visual features to a higher level representation like a set of joint positions of the body. This is a very significant step considering that low-level visual features are relatively easily obtained using current vision techniques.

2 Related Work

Previous approaches for tracking human action vary from tracking rough body position on the image plane (for example as a blob, recovering center of mass or bounding contour), to trying to find and track each body part. Tracking human bodies and hands have been the main focus of attention due to their immediate application.

One of the fundamental ideas in motion perception is the work of Johansson's moving light displays [10], where it was demonstrated that relatively little information (motion of a set of selected points on the body) is needed for humans to perform reconstruction of the body configurations. One of the first approaches for tracking walking people in real environments is due to [8]. The basic detection and registration technique commonly used is based on background segmentation, related to the work of Baumberg and Hogg [1] and Bichsel [2].

In order to find body parts using visual cues, [20] employed blob statistics and contour descriptions to roughly indicate were hands, feet, and torso were located. They needed to initialize the system with a certain body configuration in which body part identification was easy to achieve. After this, their identification relied mostly on tracking blobs. Some heuristics about body part relations, for example, *the head is at the upper most point of the segmented blob*, were used in [9]; however, this limited extensibility, also hand-crafting this rules about body position and relations requires extensive human guidance.

Model-based representations like [12, 7, 15, 17, 11, 16, 4], have also been used. The models are generally articulated bodies comprised of 2D or 3D solid primitives, sometimes accounting for self-occlusion by having an explicit body model. Multiple body configuration hypotheses were used in [5] embedded in 2D prismatic model. Most of these techniques require the use of multiple cameras, controlled viewing conditions, and/or user initialization. Also, model-based methods generally cannot recover from tracking errors in the middle of a sequence. Tracking errors may be common in real scenes where low contrast, occlusions, and changes in brightness are present. Our approach has a very low sensitivity to these effects relative to the approaches above mentioned.

The main difference in our approach with respect to model-based techniques mentioned above is that we do not try to match a body model. In our work, we do not try to match image features from frame to frame (*e.g.*,image regions, points, articulated models), as in the above set of approaches. Therefore we do not refer to our approach as tracking *per se*. Instead we are learning to map visual features to body configurations. In our approach, roughly speaking, configurations have been fully learned, no articulatory model is used. Due to this, the matching may not be as exact as the best performance of these techniques. However, it is a lot more robust and extensible, making it easier to apply to any body. It is important to mention that this way of approaching the problem may work in configurations and viewing conditions where previous tracking methods would not have any chance of giving any good results, like those shown in our experiments.

Learning based approaches include [14], where a statistical approach was taken for reconstructing the three-dimensional motions of a human figure from monocular image sequences. They used a set of motion capture examples to build a Gaussian probability model for short human motion sequences. The most influential work related to our approach is [3]. This work consisted of modeling the manifold that summarizes the given dynamical system (*e.g.*,human body motion). This manifold was modeled using a hidden Markov model and learned using a new method for entropy minimization.

Unlike these previous methods our approach does not try to model the motion characteristics of the dynamical system, but relies only of instantaneous system configuration. Even though this ignores information that can be useful for constraining the reconstruction process, it provides invariance with respect to speed and direction in which motions are performed. An important point that distinguishes our system from previous work is that we use a step of feedback matching, that transforms the reconstructed configuration back to the visual cue space to choose among a series of reconstruction hypotheses per frame. We also use a different data modeling and mapping mechanism that consists of Gaussian clusters of homogeneous configurations and a different mapping architecture based on neural networks to map body configurations from each cluster. Finally, our approach is causal while other methods are not.

3 Basic Approach

For clarity, we first very briefly enumerate every step of the proposed approach. Each of these steps will be developed with higher detail in the rest of the paper.

1) A set of motion capture sequences are obtained. This provides 3D marker positions of the given object. 2D projections of the markers are used to generate a data set Ψ of all sequences viewed from several orientations. A model of the 3D object is used to generate a set of images which are the projections of the model viewed from the same orientations. We denote the set of visual features extracted from each image Υ.

2) The data set Ψ is clustered in an unsupervised fashion to fit Gaussian distributions. This is done using the EM algorithm. In this way we obtain a set Ω of m clusters, each with roughly similar configurations.

3) For each cluster i, we train a multi-layer perceptron P_i to map visual features, in our experiments image Hu moments, from the data set Υ to the 2D marker positions Ψ.

4) Novel data is presented in the form of human silhouettes. For each frame, visual features are extracted and mapped by simulating each trained neural net.

5) The series of possible m solutions provided by each cluster is rendered in 2D space and their visual features are extracted. We then find the best match with respect to the presented data. As an optional step, consistency in time can be enforced by observing some frames ahead.

4 Learning the configuration space

4.1 Body configuration data

The source of information that will allow us to learn the object or body appearance consists of sequences of motion capture data. This provides 3D position information about the location of a set of markers. In the case of the human body data we will use, this set of markers roughly corresponds to a subset of major human body joints. This set is fixed and determined beforehand.

3D marker position can be projected into 2D marker positions at different orientations by setting the camera parameters. We use a perspective projection transformation to achieve this. We denote this set of 2D marker positions Ψ. Note that we can make this set as dense as we want by sampling at more camera orientations. Our data uses a camera located at a fixed height and distance from the center of the body.

By having knowledge of the body structure, we can render its visual appearance using computer graphics. In our case we specify the structure of the connections between markers, and use cylinders to connect them. As before, we can obtain the visual appearance of the body by setting

Figure 1: The data used for training is formed by 2D projections of 3D marker positions and their corresponding image visual features. Here we show some frames from an example sequence viewed from a given camera orientation. Training is done by sampling the set of all possible camera orientations from the same distance and height to the object.

camera parameters. We then obtain the set of visual features Υ (in our experiments image Hu moments), whose elements are in 1-1 correspondence with the elements of Ψ. Some elements of these two sets are shown in Fig. 1 where we can see 2D markers and the corresponding input image (from where visual features are extracted). We have chosen Hu moments for its easy computation, and its rotational, translational, and scale invariance.

4.2 Training to map visual features to body configurations

Given the sets Υ and Ψ (and their correspondence), we can train a neural network that maps inputs (from Υ) to outputs (from Ψ). A multi-layer perceptron with one hidden layer is employed. The explicit expression for this network is:

$$y_k = g_2(\sum_{j=0}^{l_2} w_{kj}^{(2)} g_1(\sum_{i=0}^{l_1} w_{ji}^{(1)} x_i)), \quad (1)$$

where $\mathbf{x} \in \Psi$ is the visual feature at a given instance, $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$ are each layer's synaptic weights and biases, g_1 and g_2 are a sigmoidal and linear function respectively.

This architecture was chosen because, it can approximate some non-linear mappings [13] instead of just linear ones. The architecture can also provide a 1-1 input-output correspondence, and training is relatively simple, given the data. We train the network via Levenberg-Marquardt optimization to update the weights and biases.

4.3 Clustering body configurations

It would be ideal if the mapping from Υ to Ψ were simple enough to obtain good mapping accuracy using the training procedure just explained. Unfortunately this mapping is highly ambiguous, and the visual features may not be a good descriptor of the data. Our experiments confirmed this.

For example a person facing forward would generate the same image moments as another one facing backwards. Therefore it is difficult to explicitly characterize what the weights and biases should be to create this map. Also, image moments do not encode many of the degrees of freedom for the 2D markers. Therefore, it is possible that drastically different body configurations have the same image moments descriptor.

The way we approach the above mentioned problems is by first creating clusters of statistically homogeneous data points in the 2D projected marker space. We try to generate clusters that can be described by a Gaussian probability distribution. This is an unsupervised clustering task, for this we use the Expectation Maximization algorithm [6].

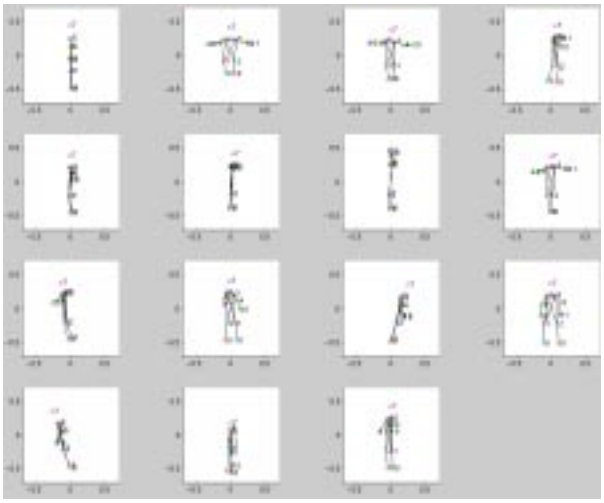

Figure 2: The cluster means obtained after performing unsupervised clustering of the data points. Each cluster's distribution is approximated by a Gaussian probability distribution. Note that opposite configurations () are clustered separately. For example, one can see that there is a cluster for the figure facing forward, and another one backward. This separation is important because visual features alone cannot resolve this ambiguity. Therefore, the complexity of the mapping can be reduced if clusters are trained separately.

Let's denote $\Theta_i = (\mu_i, \Sigma_i)$ to be the learned distribution parameters for cluster i. For each data point $\mathbf{x} \in \Psi$, we can assign it to a cluster, by just finding the ML (maximun-likelihood) estimate.

$$i = \arg\max_j(P(\Theta_j|\mathbf{x})) = \arg\max_j(N(\mathbf{x}, \mu_j, \Sigma_j)), \quad (2)$$

where i is the label of the cluster to which we assigned this data point $\mathbf{x}$. Fig. 2 illustrates this idea, it shows the mean configurations for a set of 15 clusters found by this method.

Once data points are divided into clusters, a neural network is trained for each cluster as described in Sec. 4.2. This results in a set $\Omega = \{P_1, P_2, ..., P_m\}$ of m neural networks, each trained to particular body configurations.

This architecture is based on the idea that by partitioning (via clustering) the body configuration space into ho-

mogeneous regions, we can learn a more specialized map from visual feature space. This reduces the ambiguities mentioned above. For example, in Fig. 2 we can see that there are mean configurations facing forward and backward. This indicates that it is very likely that a different cluster will be used to map visual features for each of these orientations. This should therefore reduce the ambiguities in the mapping and make it simpler to learn.

5 Synthesizing body configurations

The practical goal of this work is to map raw visual data of human silhouettes to a structured set of body feature location (in our case joint positions).

When novel data $\mathbf{x}$ is presented, we first simulate each of the neural networks in Ω, yielding a set of hypothesis of body configurations $\mathbf{T} = \{\mathbf{y}_k\}$, with $k = 1...m$, and m is the number of clusters.

The question is: how to choose from this set of hypotheses? We approach this problem by creating another mapping P_b using a multi-layer perceptron, in this case trained to map from points in marker space to points in visual feature space. This can be done using the sets Υ and Ψ. Because this mapping uses data rendered with knowledge of 3D information, it is very likely to have accuracy advantages over the simpler transformation, which renders the 2D markers to produce a 2D image and then it finds its visual features. Recall that the visual features where produced by a 3D body, not by a 2D one.

The reason for this mapping is justified as follows. Once we obtain the set $\mathbf{Y}$ of hypotheses about the body configuration by observing the visual feature $\mathbf{x}$, we can map each element $\mathbf{y}_k$ ($k = 1..m$) back to visual feature space, and obtain representations $\tilde{\mathbf{x}}_k$. The most accurate hypothesis is found by minimizing:

$$i = \arg\min_j (R(\mathbf{y}_j) - \mathbf{x}_j)^\top \Sigma_\Upsilon^{-1} (R(\mathbf{y}_j) - \mathbf{x}_j), \qquad (3)$$

where Σ_Υ is the covariance matrix of the set Υ, R is a function that maps marker positions to visual features, j varies over the set of hypotheses, and i is the neural network label that best matched the visual feature observed.

As a further refinement step, because neighboring frames are generally from similar configurations, we have obtained slightly better performance if consistency in time is enforced. Therefore, after we obtain the best network to use for a given frame, if this network differs from that in the previous frame, we wait for more frames to arrive (generally 2 or 3) to decide whether to use this new network. If within this window, the new frames are consistent with the change, the new network is used, if not the previous network is used instead. This was found to be useful in detecting spurious individual reconstructed frames. The use of motion information in the system is an issue that requires further research work.

6 Experiments

In order to evaluate the performance of our approach, we conducted experiments in which we had knowledge of the *best* reconstruction. Using the data sets Ψ and Υ, we performed clustering and training taking out the sequence with the specific orientation that would be used for testing. We also took out its neighboring views, the opposite view and its neighbors. View orientations were sampled every $\pi/16$ radians, for a total of 32 orientations. The training data set consisted of five sequences with an average of about 200 frames each, sampled at the 32 orientations above mentioned. The 3D motion-capture data was obtained from http://www.biovision.com. This data consisted of position information of 37 markers, from which we chose a subset of 11, considered by us the most informative ones.

Fig. 3 shows the reconstruction obtained by our approach when images of a *dance* sequence were shown. The view angles used were $4\pi/32$ and $7\pi/32$ radians respectively. The agreement between reconstruction and ground-truth is easy to perceive for all sequences. Note that for self-occluding configurations, like the second frame of Fig. 3, reconstruction is more difficult, but still the estimate is very close. This is mainly due to the inadequacy of the feature and image representation to separate certain configurations that are different in the marker space.

Another reconstructed sequence (*destroy*)is shown in Fig. 4, obtaining similar results as before. Note that the body is also turning around its axis. As can be seen in the figure, there are very challenging configurations and orientations, which are correctly reconstructed by our approach. Sequences are shown for 0 and $7\pi/32$ radians.

In order to formally characterize the performance of our system, using the training and testing procedure described above, we measured the average marker error (measured as the distance between reconstructed and ground-truth marker). After testing all the sequences, the mean and variance marker displacement was 0.0289 and 0.000422 units respectively. As a point or reference, the height of the figure was approximately 1.4 units in average.

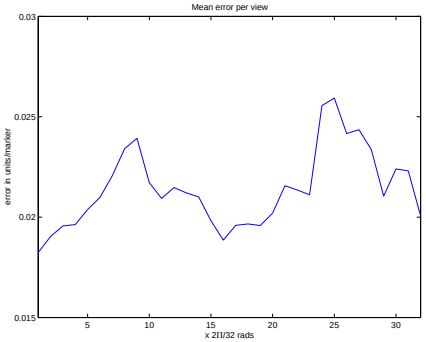

Figure 5: Measure of the mean marker error per view angle. Figures were aligned to always face forward for the 0 radians view angle. Views are taken every $\pi/32$ radians starting at 0 radians. Note that the error is bigger for orientations close to $\pi/2$ and $3\pi/2$ radians.

We also measured the average error marker per body orientation. For this we rotated the 3D figures so that their orientation corresponds to the orientation tested. Recall that in the original sequences, bodies are not always facing a fixed point. Angles are sampled every $\pi/32$ radians

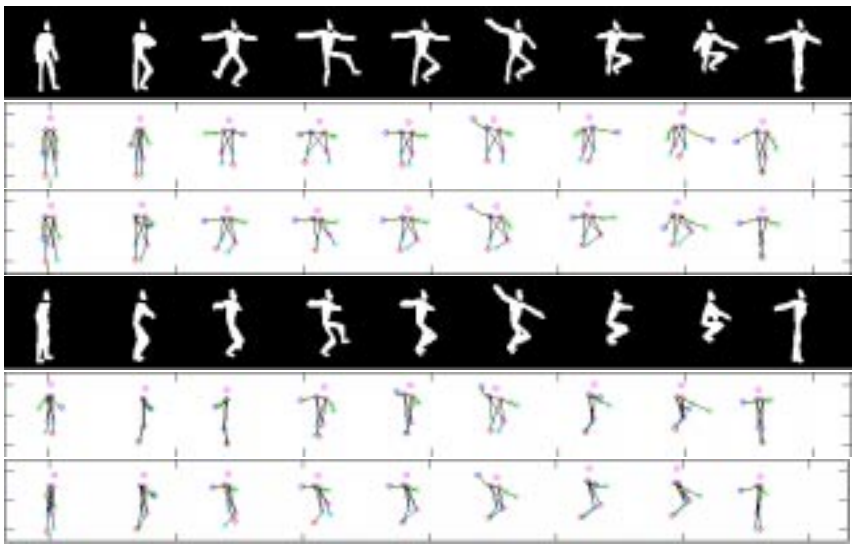

Figure 3: Example reconstruction of one of the *dance* sequence. Each set (3 rows each) consists of input images, reconstruction, and ground-truth. Results are shown every 25th frame. View angles are $4\pi/32$, and $7\pi/32$ radians. The obtained reconstruction visually agrees with the ground-truth output for all views.

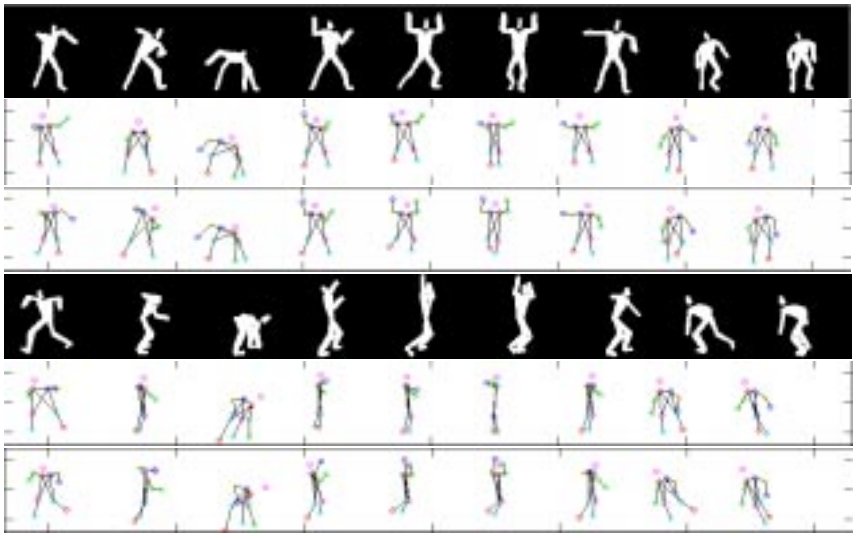

Figure 4: Example reconstruction of the *destroy* sequence. Each set (3 rows each) consists of input images, reconstruction, and ground-truth. Results are shown every 25th frame. View angles are 0 and $7\pi/32$ radians. The obtained reconstruction visually agrees with the perfect output for all views. Note that this sequence has challenging configurations, body orientation is also recovered correctly.

starting at 0 radians, which corresponds to the person always facing to the camera. Note that the error is bigger for orientations closer to $\pi/2$ and $3\pi/2$ radians. This intuitively agrees with the notion that at those angles, there is less visibility of the body parts, therefore making it harder to use visual features in distinguishing between different configurations (*i.e.*,different configurations are closer in visual feature space). This performance is very promising considering the complexity of the task and the simplicity of the approach.

In the next example, in Fig. 6 we tested the system against real segmented visual data, obtained from observing and tracking people walking in an outdoor environment. We chose a walking sequence at an angle between $\pi/4$ and $\pi/2$ radians. Note that even though the characteristics of the segmented body differ from the ones used for training, good performance is achieved. Body orientation appears to be recovered correctly, even though according to our performance evaluations, this orientation is among the hardest to recover. We have tested several actions (waving, crouching-down, leaning over, walking and running) and the performance is comparable. Even though we expect to perform a wider variety of experiments, this experiment begins to demonstrate the promise of our approach in dealing with real visual data. Of course, the closer the real body visual features are to the training data, the better the expected reconstruction. Note that the results shown were obtained by training just once, there was no need to choose

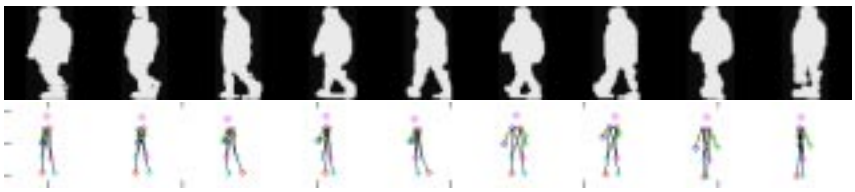

Figure 6: Example reconstruction sequence from real video data of a person walking. Input images and reconstruction are shown for nine frames. Results are shown every 15th frame. The obtained reconstruction visually agrees with the input.

among the best of trained models, indicating the potential of the approach.

7 Discussion

We have presented a novel technique that allows the reconstruction of human body configuration from raw low-level visual features. The approach is both simple and powerful. Due to its generality, the approach can be used for learning mappings for other non-rigid or articulated objects.

Human pose reconstruction is a particularly hard problem because this mapping is highly ambiguous. We have obtained excellent results even using a very simple set of image features, such as image moments. Choosing the best subset of image features from a given set is by itself a challenging problem, and a topic of on-going research.

Our ideas are different from tracking approaches in that we do not try to match body parts using image patches from frame to frame. Instead we follow a statistical approach. By learning a subspace of body configurations, the system can constrain the direct mapping from visual features to body configuration. Because of the complexity of the mapping, we clustered the space of 2D body configurations into approximately homogeneous configurations, showing improved results.

The implemented algorithm for reconstruction runs in linear time $O(M)$ with respect to the number of clusters M. It scales linearly for sequences, for a sequence of length N, the complexity is $O(NM)$. The method is by itself causal, but performance improved slightly when looking 2 or 3 frames ahead. Just as an interesting point, it is believed that human perception is delayed several milliseconds [19].

The system was tested in recovering the pose for both synthetic and real visual data. The synthetic data was used for measuring the performance of the approach, real data showed its applicability. Reconstructing body pose is an important step towards advanced automatic human motion analysis. The results of the experiments are encouraging considering the complexity of the task and previous results.

Acknowledgments

Some of the ideas here presented were developed while the first author was at MERL. Thanks to Matt Brand and Aaron Hertzmann for sharing excellent and insightful discussions. This work was supported in part through Office of Naval Research Young Investigator Award N00014-96-1-0661, and National Science Foundation grants IIS-9624168 and EIA-9623865.

References

[1]A. Baumberg and D. Hogg. Learning flexible models from image sequences. In *ECCV*, 1994.

[2]M. Bichsel. Segmenting simply connected moving objects in a static scene. *PAMI*, 16 (11):1138-1142, 1994.

[3]M. Brand. Shadow puppetry. In *ICCV*, 1999.

[4]C. Bregler. Tracking people with twists and exponential maps. In *CVPR*, 1998.

[5]T.J. Cham and J. Rehg. A multiple hypothesis approach to figure tracking. In *CVPR*, 1999.

[6]A. Dempster, N. Laird, and D. Rubin. Maximum likelihood estimation from incomplete data. *Journal of the Royal Statistical Society (B)*, 39(1), 1977.

[7]D. Gavrila and L. Davis. Tracking of humans in action: a 3-d model-based approac. In *Proc. ARPA Image Understanding Workshop, Palm Springs*, 1996.

[8]D. Hogg. *Interpreting Images of a Known Moving Object*. PhD thesis, University of Sussex, 1984.

[9]L. Davis I. Haritaouglu, D. Harwood. Ghost: A human body part labeling system using silhouettes. In *Intl. Conf. Pattern Recognition*, 1998.

[10]G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception and Psychophysics*, 14(2): 210-211, 1973.

[11]S. Ju, M. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated image motion. In *Proc. Gesture Recognition*, 1996.

[12]I. Kakadiaris, D. Metaxas, and R. Bajcsy. Active part-decomposition, shape and motion estimation of articulated objects: A physics-based approach. In *CVPR*, 1994.

[13]A. Lapedes and R Farber. How neural nets work. *Neural Information Processing Systems*, 442-456, 1988.

[14]M. Leventon and W. Freeman. Bayesian estimation of 3-d human motion. Technical Report 98-06, Mitsubishi Electric Research Labs, 1998.

[15]A. Pentland and B. Horowitz. Recovery of non-rigid motion and structure. *PAMI*, 13(7):730–742, 1991.

[16]J. M. Regh and T. Kanade. Model-based tracking of self-occluding articulated objects. In *ICCV*, 1995.

[17]K. Rohr. Towards model-based recognition of human movements in image sequences. *CVGIP:IU*, 59(1):94-115, 1994.

[18]R. Rosales and S. Sclaroff. 3d trajectory recovery for tracking multiple objects and trajectory guided recognition of actions. In *CVPR*, 1999.

[19]L. Spillmann and J. Werner. *Visual Perception*. Academic Press, 1990.

[20]C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland. Pfinder: Real time tracking of the human body. *PAMI*, 19(7):780-785, 1997.