Automatic Acquisition and Initialization of Articulated Models

N. Krahnstoever* M. Yeasin R. Sharma

Department of Computer Science and Engineering

The Pennsylvania State University

220 Pond Lab, University Park, PA 16802

E-Mail: {krahnsto,yeasin,rsharma}@cse.psu.edu

Abstract

Tracking, classification and visual analysis of articulated motion is challenging due to the difficulties involved in separating noise and variabilities caused by appearance, size and view point fluctuations from task-relevant variations. By incorporating powerful domain knowledge, model based approaches are able to overcome these problem to a great extent and are actively explored by many researchers. However, model acquisition, initialization and adaptation are still relatively under-investigated problems, especially for the case of single camera systems.

In this paper, we address the problem of automatic acquisition and initialization of articulated models from monocular video without any prior knowledge of shape and kinematic structure. The framework is applied in a human computer interaction context where articulated shape models have to be acquired from unknown users for subsequent limb tracking. Bayesian motion segmentation is used to extract and initialize articulated models from visual data from the ground up. Image sequences are decomposed into rigid components that can undergo parametric motion. The relative motion of

*Corresponding author. E-Mail: krahnsto@cse.psu.edu, Phone: +1 (814) 865-2729, Fax: +1 (814) 865-3176.

these components is used to obtain joint information. The resulting components are assembled into an articulated kinematic model which is then used for visual tracking eliminating the need for manual initialization or adaptation. The efficacy of the method is demonstrated on synthetic as well as natural image sequences. The accuracy of the joint estimation stage is verified on ground truth data.

1. Introduction

Important capabilities of vision based human-computer interaction systems are the detection, capture, analysis and synthesis of human motion. However, the processing of human-motion is extremely challenging due to (i) non-rigid motion patterns caused by the inherent nature of the articulated human body and clothes, (ii) self-occlusion and (iii) lack of visual texture. Furthermore, the extraction of features that are suitable for view-invariant recognition and classification (e.g., hand-gestures or human actions and activity in general) is challenging due to the strong view-point dependent variabilities of the visual motion patterns. The use of explicit articulated models is promising for overcoming these challenges because it allows to directly encode much of the available domain knowledge and potentially offers a wider degree of generality and task independence than existing approaches.

Remaining challenges that model based approaches face are model acquisition, initialization and adaptation [1]. Model acquisition is the process of constructing the articulated model that encodes the information about the limbs and the interconnecting joints. Articulated models come in many different flavors with varying number of links and joints and are commonly hand crafted. Since the size and shape of people varies across the population, it is usually not possible to develop universal models. Models, especially the limb shape parameters, have to be adapted to the dimensions and the appearance of the target. Finally, most model based tracking approaches assume that the sequence. This problem of model initialization is commonly reported to be performed manually by the user. The use of model based motion capture systems in many domains (e.g., surveillance, human-computer interaction, automatic video indexing), will in general only become feasible, once the above challenges have been tackled.

This work is motivated by and aimed at the domain of human computer interaction and

gesture recognition applications [2, 3, 4] where the goal is to robustly track a user over time without any manual initialization. We propose to eliminate the need for initialization and adaptation by automatically building articulated models from visual data directly. Our approach assembles articulated models from monocular video from the ground up assuming only the "concept" of articulated motion as prior knowledge. We assume that the "world" consists of rigid segments that are potentially connected by joints and leave it to the algorithm to extract segment and joint information automatically from an image sequence. More specifically, we use a parametric motion segmentation approach [5, 6, 7] to simultaneously decompose a set of images into rigid segments, together with their corresponding motion parameters. The motion models of the layers are subsequently examined to infer joint locations. The combination of the extracted segments, their motion parameters and joint locations constitutes a complete articulated model with joints, links and appearance information. We show how the acquired and initialized articulated models can be used for tracking and motion capture. Furthermore, we quantitatively evaluate the accuracy of the approach based on models extracted from synthetic image sequences generated with professional character animation tools for which precise knowledge about joint locations is available.

This paper is organized as follows: We review related work in Section 2. Following this, we present our approach to extract the rigid components of the input sequence that form the link candidates in Section 3. Section 4 describes the model extraction stage which is responsible for detecting and locating joints and for inferring which of the extracted motion segments are part of the observed target. To evaluate the extracted model we implemented a model based tracking algorithm, which is briefly described in Section 5. Experiments on real and synthetic data are presented in Section 6 followed by a discussion of these results in Section 7. Finally, Section 8 concludes the paper.

2. Related Work

2.1. Analysis of Articulated Motion

With respect to the visual analysis of articulated motion, much research has been conducted on the analysis of feature point models where the visual information is reduced to a set of points

attached to the rigid body. This type of data can arise from Moving Light Displays (MLDs), passive or active markers or extracted from image sequences using feature trackers.

Early work by Rashid [8] presented a comprehensive algorithm for analyzing MLDs. Point features where clustered into objects using a minimum spanning tree (MST) approach together with a cut criterion for splitting the resulting tree into clusters. The underlying skeletal structure of the MLD groups was obtained by calculating MST on each group. Rashid stressed the importance of velocity information in obtaining robust estimates of skeletal structure.

Holt et. al [9] address the problem of recovering the 3D motion of articulated objects from observed time varying 2D joint locations. Their approach constrained the allowable observations such as the assumption of planar motion of the objects arms with respect to a central torso.

Using magnetic motion capture data obtained with magnetic sensors, O'Brien et al. [10] reconstructed the skeletal model of articulated objects and humans. Their approach is based on available time varying 3D marker coordinate systems. These systems are examined for joint constraints and a MST is employed to reconstruct the articulated structure. While their method assumes knowledge about the 3D link coordinate systems, the basis of their approach is also applicable to projected motion data.

A motion modeling approach not based on the articulated structure of the human body was developed by Song et al. [11, 12]. The front view of the human body is modeled through a set of point feature tracks whose locations are modeled through conditional probability densities. The densities are learned from training data.

2.2. Human Shape Acquisition

The acquisition of precise human body shape models has so far mostly been investigated for situation in which multiple camera views are available. In [13] human body models without any prior structural assumptions are acquired using a large number of camera views in a customized laboratory environment. Acquisition systems that require the views of at least three cameras to perform the modeling process have been presented by [14] and [15]. The latter work uses a generic body model onto which the appearance of a persons view is mapped.

The work presented in [16] utilizes a stereo setup and uses a flexible human shape model to adapt to the shape of a user in the view of the camera. A manual initialization of the shape model is necessary for bootstrapping the procedure. In general all these efforts are not suitable for environments where only a single camera view is available, which is especially the case for low-cost HCI applications. Furthermore, a flexible acquisition procedure should not be restricted to an a priori given articulated structure.

Ioffe et. al [17] uses tree structured probabilistic models for modeling human motion from monocular video. While elegant and not based on any structural assumptions, the approach is based on the ability extract candidate body parts from static images and only utilizes weak motion models.

Similar to Ioffe's goals, the work presented in this paper allows to acquire articulated models consisting of planar image patches connected by joints and thus falls neither into the category of MLDs nor into the class of algorithms that acquire "inflated" three dimensional models. The obtained models resemble the cardboard type articulated models that have been shown to provide utility in many applications [18, 19].

2.3. Motion Segmentation

One significant portion of this work deals with the motion segmentation of image sequences for extracting the piecewise rigid components of the articulated objects. Recent years have seen a great interest in layered motion segmentation algorithms [6, 5, 20, 21, 7]. These algorithms address the problem of segmentation and flow estimation in a unified framework to overcome some of the main problems of either method alone. While the early work of Wang and Adelson [7] and subsequent improvements [20, 21] approached the problem using clustering, the problem has since been formulated in an elegant Bayesian framework based on an expectation-maximization (EM) [22].

2.4. Model Based Tracking

The approach of *model based human tracking* has been pioneered by O'Rourke and Badler [23]. In the context of human (or general articulated) motion tracking, the target is modeled

as a collection of segments connected by joints or springs. The number of links and joints and associated parameters used for articulated models vary widely across the literature [24].

Ju et al. [18] approximate humans with *cardboard models*, which are basically 2D models specialized at modeling humans seen from the side or front. Each segment of the model is described by a planar patch that can undergo planar projective motion. The motion of the patches is determined through an energy function that uses the brightness constancy constraint equation and spring like forces between connected patches. Tracking is achieved by minimizing the energy function using gradient descent in a hierarchical framework. Improvements utilize joints [25] and handle occlusion [26].

Pavlovic et al. [27] address the problem of learning dynamics from training data in a Bayesian framework. They also employ a 2D model but use scaled prismatics [28] that are able to handle 3D foreshortening effects and avoid singularities common in 3D kinematic modeling approaches.

Gavrila and Davis [29] developed a four camera full-body tracking system using a 3D model of tapered super-quadrics. The system was able to successfully track two people dancing close together in the presence of strong occlusion. Pose estimation was performed using search space decomposition and best-first search.

The model-based approach to arm tracking is particularly promising in HCI application and has been addressed in [30, 31, 32, 33]. In all cases, two link models were used.

Instead of energy minimization or variational frameworks, the use of sequential monte carlo methods [34] is gaining popularity. Sidenbladh et. al [35] perform 3D reconstruction human motion observed with a single camera using models consisting of ten cylinders under perspective projection connected by joints parameterized by 25 values. Tracking was performed using a particle filtering approach [36]. Appearance information of the cylinders is adapted incrementally from the image sequences. The authors encouraged the use of more persistent appearance models. Furthermore, in a recent paper [37], Sidenbladh showed that the performance of the approach could be improved further by learning the parameters for the edge and ridge filters used in the likelihood model. These aspects encourage finely tuned

models that are learned or acquired from data.

Deutscher et al. [38] also use a particle filtering approach to human motion tracking. They point out that the high dimensionality of the articulated models are the main problem when trying to recover articulated pose over time. Particle trackers in general need a number of particles that is exponential in the dimensionality of the problem. The method presented by the authors reduces this number by searching for the global maximum of a general weighting function $w(Z_k, X)$ in an annealing type approach.

Bregler and Malik [39] developed a articulated motion capture framework that parameterizes the kinematic chain of the human body in an exponential twist formulation pioneered in robotics [40]. In addition, the authors do not follow a synthesize and match approach but rather developed a variational approach that relies on matching the appearance of limbs with the image content. The differential formulation is ultimately based on the brightness constancy assumption and linear approximations of the twists. Their results are remarkable as they were able to report good tracking results on very noisy image sequences. We believe that most of the performance of his approach stems from an explicit handling of depth information (i.e., occlusion) and the use of explicit shape *and* appearance information of the limbs, encouraging to learn such models from data.

Covell et. al [41], presented extensions to the twist and exponential map tracking framework of [39]. Other research that utilizes the twist-formulation was presented in [42].

Drummond and Cipolla [43] also use twists but performs human body tracking using kinematic trees with links defined as the contours (conics) of truncated quadrics. Link motions are associated with probability densities through a simple Taylor expansion of the match function. Articulated constraints between links are enforced by propagating likelihood densities along the kinematic chain.

3. Link Extraction

The acquisition of articulated models from visual data involves three main steps: The detection and extraction of the links that are assumed to give rise to piecewise rigid motion patterns in

the video, the detection and localization of joint constraints and joint centers between the links and the final assembly of the model.

3.1. Motion Segmentation

The link extraction is achieved by performing motion segmentation on a sequence of video images. Motion segmentation algorithms take two images as input and perform a segmentation into non-overlapping regions that move according to independent parameterized motion models. Every pixel in the reference image is assigned to one of K layers $L_i, i = 1, \ldots, K$ or designated as an outlier. For every layer, the motion parameters are estimated. The extraction of layer segmentation and motion parameters is performed using the expectation maximization (EM) algorithm [22].

In this approach, motion segmentation is performed iteratively in two stages until convergence : In the expectation stage, the motion parameters are assumed to be known and the layer assignments are estimated for every pixel. In the maximization stage, the assignments are assumed known and the motion parameters are estimated. The EM approach maximizes the overall likelihood of layer assignments and motion parameters and leads to very good results if the algorithm starts with reasonable initial values (cf. Sect. 3.2).

Our approach to performing the motion segmentation is based the works of [5, 6] which are both two-frame algorithms. In our domain, the following problem arises when performing motion segmentation on two frames: in order to obtain a decomposition of the articulated target into segments that correspond to limbs based on motion information, it is necessary that each segment undergoes motion that distinguishes itself from the motion of all other segments as far as motion parameters are concerned. If two limbs perform the same motion, they cannot be distinguished from each other and will be viewed as one part. Hence we augmented the motion segmentation in multiple directions: First, to order to increase the chance of observing distinguishable motion patterns for pairs of segments, our algorithm performs the motion segmentation on one frame, while drawing motion information from several frames. Furthermore, we incorporate appearance and shape information into the estimation procedure.

Since the result of the EM algorithm tends to be only as good as the initial estimate, the

initialization approach is an important part of this work. For initialization we use a sparse flow clustering method to obtain initial estimates of the motion parameters. Our method is outlined as follows (see Figure 1): The motion segmentation is performed for a reference frame I_0 based on a set of N_F previous and subsequent images $\mathbf{I}_{N_F} = \mathbf{I}^+_{N_F} \cup \mathbf{I}^-_{N_F}$ with $\mathbf{I}^\pm_{N_F} = \{I_{\pm 1}, \ldots, I_{\pm N_F}\}$. The value of N_F is usually chosen to be around $2-6$. The algorithm estimates motion parameters θ_{if} that map the i^{th} layer L_i from image I_f to I_0 and layer assignment probabilities $\lambda_i(\mathbf{x})$ that denote the probability of pixel $\mathbf{x}$ in I_0 belonging to the i^{th} layer L_i. In order to handle effects caused by occlusion, the motion segmentation is performed separately in forward and backward direction and the results combined in a final stage. We will outline the forward case here. Fig. 1

With Bayes rule we have (cf. [5]):

$$\begin{aligned}\lambda_{\text{i}}(\mathbf{x}) &= P(\mathbf{x} \in L_{\text{i}} | I_0(\mathbf{x}), \mathbf{I}^+_{N_F}, \mathbf{\Theta}_{\text{i}}, \Psi_i) \\ &= cP(I_0(\mathbf{x}) | \mathbf{x} \in L_{\text{i}}, \mathbf{I}^+_{N_F}, \mathbf{\Theta}_{\text{i}}, \Psi_i) P(\mathbf{x} \in L_{\text{i}} | \mathbf{I}^+_{N_F}, \mathbf{\Theta}_{\text{i}}), \end{aligned} \tag{1}$$

with $\mathbf{\Theta}_i = \{\theta_{if}, f = 1, \ldots, N_F\}$ and Ψ_i denoting additional shape and color parameters to be specified shortly. The first term on the right hand side of Eq. (1) expresses the likelihood of the observed image given the current segmentation and motion parameters. In our motion segmentation implementation, this term draws its information from three sources

$$\begin{aligned}P(I_0(\mathbf{x}) | \mathbf{x} \in L_i, \mathbf{I}^+_{N_F} \mathbf{\Theta}_i, \Psi_i) &= P_{\text{r}}(I_0(\mathbf{x}) | \mathbf{x} \in L_i, \mathbf{I}^+_{N_F}, \mathbf{\Theta}_i) \\ &\quad \cdot P_{\text{s}}(\mathbf{x} | \mathbf{x} \in L_i, \Psi_i) \\ &\quad \cdot P_{\text{c}}(I_0(\mathbf{x}) | \mathbf{x} \in L_i, \Psi_i) \end{aligned} \tag{2}$$

where $P_{\text{r}}(.)$ models the residuals arising from the match between I_0 and the following images given the motion parameters, $P_{\text{s}}(.)$ expresses the conformance to the shape and $P_{\text{c}}(.)$ to the color model of the i^{th} layer. The residual term $P_{\text{r}}(.)$ on the right hand side of Eq. (2) is assumed to be normally distributed in the residuals originating from the match of the layers at their location in I_0 and in the frames $\mathbf{I}^+_{N_F}$

$$P_{\text{r}}(I_0(\mathbf{x}) | \mathbf{x} \in L_i, \mathbf{I}^+_{N_F} \mathbf{\Theta}_i) = \prod_{f=1}^{N_F} \mathcal{N}(r_{if}(\mathbf{x}); \sigma_i), \tag{3}$$

with residuals

$$r_{if}(\mathbf{x}) = I_0(\mathbf{x}) - I_f(\mathbf{P}(\mathbf{x}; \theta_{if})). \tag{4}$$

The function $\mathbf{P}(\mathbf{x}; \theta)$ denotes a warp function that maps pixels from images in $\mathbf{I}^+_{N_F}$ to their location in the reference frame. The matching residuals can be viewed as the errors associated with a backward prediction of frame I_0 by subsequent frames. The second term $P_{\mathrm{s}}(.)$ in (2) assumes that the pixels in a layer are normally distributed around a center location μ_i with empirical shape covariance matrix Σ_i

$$P_{\mathrm{s}}(\mathbf{x}|\mathbf{x} \in L_i, \mathbf{I}^+_{N_F}\mathbf{\Theta}_i) = \frac{1}{2\pi|\Sigma_i|} e^{-\frac{1}{2}(\mathbf{x}-\mu_i)^T \Sigma_i^{-1}(\mathbf{x}-\mu_i)}. \tag{5}$$

This effectively leads to a blob like clustering of pixels and helps to obtain compact layer supports for the link shapes. It also helps to resolve ambiguous assignments such as pixels from untextured regions for which any motion model would locally describe the visual data correctly.

The third term $P_{\mathrm{c}}(.)$ in Eq. (2) expresses the conformance of a pixel in the i^{th} layer to other pixels in this layer. We assume that pixel values are normally distributed in RGB color space according to

$$P_{\mathrm{c}}(I_0(\mathbf{x})|\mathbf{x} \in L_i, \mathbf{I}^+_{N_F}\mathbf{\Theta}_i) = \frac{1}{(2\pi)^{\frac{3}{2}}|\Sigma_i^{\mathrm{C}}|} e^{-\frac{1}{2}(I_0(\mathbf{x})-\mu_i^{\mathrm{C}})^T \Sigma_i^{\mathrm{C}\,-1}(I_0(\mathbf{x})-\mu_i^{\mathrm{C}})}. \tag{6}$$

This term helps to improve the assignment of pixels to layers especially at the boundaries of layer regions and in untextured areas. The parameters of these residual, shape and color models,

$$\Psi = (\Psi_1, \ldots, \Psi_K) \text{ with } \Psi_i = (\sigma_i, \mu_i, \Sigma_i, \mu_i^{\mathrm{C}}, \Sigma_i^{\mathrm{C}}) \tag{7}$$

are estimated after the maximization stage of the EM algorithm before the layer assignment calculation.

For the motion model a six parameter affine 2D transform is used

$$\mathbf{P}(\mathbf{x}; \theta) = \mathbf{A}x + \mathbf{t}, \tag{8}$$

with parameters

$$\theta = (A_{11}, A_{12}, A_{21}, A_{22}, t_1, t_2). \tag{9}$$

The second term in Eq. (1) is the assignment prior which can be chosen to be independent of $\mathbf{x}$ (as in [6]) or used to impose smoothness on the layers (as in [5]). We follow the latter option and use a MRF prior [5] that effectively enforces a spatial coherence in the layer assignments and leads to smoother results .

At each **E-step**, the calculation of (1) and the MRF prior has to be iterated to obtain the posterior layer assignment estimates:

The **M-step** assumes known $\lambda_i(\mathbf{x})$ and minimizes the prediction error

$$h(\mathbf{\Theta}) = \sum_{f,i,x} \lambda_i(\mathbf{x}) \frac{r_{if}(\mathbf{x})^2}{\sigma_i^2}. \tag{10}$$

This step can be interpreted as a simple simultaneous registration of image I_0 to the N_F frames with the support restricted according to the layer assignments. Equation (10) separates into a sum of independent terms. The gradient and Hessian of Eq. (10) can be calculated easily [6] and we minimize $h(\mathbf{\Theta})$ using Gauss Newton optimization with line search [44]. The EM iteration has to be repeated until convergence. For details about how to incorporate the estimation of the prior parameters into the EM framework see ref. [5].

3.2. Initialization

The EM algorithm is guaranteed to maximize the likelihood of the solution but can get stuck in local maxima. A good initialization of the motion parameters is hence crucial for the success of the algorithm. We initialize the procedure by first performing sparse motion estimation [45, 46] across a time interval of images with indices $[N_1, N_2]$ that includes the images from which the motion segmentation is performed (i.e., $N_1 \leq -N_F$ and $N_2 \geq N_F$). The sparse motion estimation yields a set of N_T feature tracks $\mathbf{y}_i^t$ with $i \in 1, \ldots, N_T$ and $t \in [N_1, N_2]$. Each feature track is assumed to move with one of the K regions in the image. Motion parameter estimates can thus be obtained from the motion of the feature tracks if the assignment of features to regions is known. This assignment can be obtained through simple K-means

feature track clustering with a track distance function defined as follows:

$$d(\mathbf{y}_i, \mathbf{y}_j) = \sum_{t=N_1}^{N_2} (\Delta(\mathbf{y}_i^t, \mathbf{y}_j^t) - \bar{\Delta}_{ij})^2 + \alpha \sum_{t=N_1}^{N_2-1} ||\mathbf{v}_i^t - \mathbf{v}_j^t||^2 \tag{11}$$

with $\Delta(\mathbf{a}, \mathbf{b}) = ||\mathbf{a}-\mathbf{b}||$, $\bar{\Delta}_{ij}$ the mean of $\Delta(\mathbf{y}_i^t, \mathbf{y}_j^t)$ over all t and $\mathbf{v}_i^t = \mathbf{y}_i^{t+1} - \mathbf{y}_i^t$. This distance function expresses the fact that two feature tracks are considered to move with the same layer if their relative distance varies little across time and their velocity is similar. After the K-means clustering, affine motion models are estimated from the grouped feature tracks using a standard least squares method. These estimates are used to bootstrap the EM procedure for the motion segmentation. In practice, the feature track initialization method leads to very good initial estimates reducing the burden on the motion estimation step in the motion segmentation stage considerably leading to a rapid convergence.

3.3. Refinement of Motion Estimates

Because the motion segmentation procedure does not utilize or estimate any depth ordering of the regions, artifacts can occur in assignments at the layer boundaries where pixels in the image become occluded in subsequent frames. The occurrence of these ambiguities increases with the displacement of the layers and the number of frames used for the estimation but occurs only in the direction of the layer movement. It can hence be canceled out by performing motion segmentation in both forward and backward direction with respect to the reference frame. The final layers are then given by the intersection between the forward and backward estimated layers which improves the quality of the support regions substantially. Figure 1 summarizes the flow of information during the initialization and segmentation procedure.

The final link regions are obtained by labeling the connected components of the layer assignment mask and subsequent extraction of the largest connected component. This obtains a single connected region of support for each link. A tight bounding box is calculated for each resulting support regions and the image content extracted together with its alpha map. This image information constitutes the size and appearance information for each link. With this information, the motion estimation stage of the motion segmentation algorithm is restarted with the layer assignments fixed according to the thus extracted link regions. The number of images for

which the motion parameters of the layers are estimated is increased to an interval $[1, \ldots, N_J]$ in order to obtain extended estimates of the motion of the links in order to improve the extraction of joint information in the next stage. For time instances $(N_{F+1}, \ldots, \min\{N_J, N_2\}]$, the motion estimates from the feature track initialization stage can again be used to initialize the motion estimation procedure.

4. Model Extraction

The motion segmentation stage decomposes the reference image I_0 into a set of connected rigidly moving regions and yields the parameters of the transformations that maps these regions to the images $\{I_1, \ldots, I_{N_J}\}$. Each individual region may or may not be a link of the target subject. The goal of the model extraction stage is to decide which regions in the image are components of the model and to detect and locate joint connections between these components. Each of the regions extracted in the motion segmentation stage is considered to be a potential link of the articulated model.

In the following, we denote with $\mathbf{T}_i$ the transformation that maps a point $\mathbf{x}_w$ from world coordinates to the i^{th} link coordinate system $\mathbf{x}_i = \mathbf{T}_i(\mathbf{x}_w)$ at time $t = 0$. With $\mathbf{P}_i^f$ we denote the transformation that maps a world coordinate at time $t = 0$ to the world coordinate system at time t under the assumption that it moved according to the motion of the i^{th} link, $\mathbf{x}_{wi}^t = \mathbf{P}_i^t(\mathbf{x}_w)$.

In general, if the relative pose of two coordinate systems C_i, C_j is constrained by the existence of a rotational joint between them, there must exist two points $\mathbf{x_i} \in C_1$ and $\mathbf{x_j} \in C_2$ that always map to the same world coordinates, $\mathbf{P}_i^t(T_i^{-1}(\mathbf{x}_i)) = \mathbf{P}_j^t(T_j^{-1}(\mathbf{x}_j))$, for all t. The points x_i and x_j are the link coordinates of the joint center. In a strict sense, the converse is not true. The existence of two such points does not guarantee the existence of a joint, especially if the motion of these objects is only observed in an image plane projection.

However, if two such points exist between two objects that do move non-uniformly with respect to each other over extended periods of time, it is reasonable to assume that this indi-

cates the existence of a joint. More specifically, if the average *link coincidence*

$$d_{ij}^2 = \min_{(\mathbf{x}_i,\mathbf{x}_j)} d(\mathbf{x}_i,\mathbf{x}_j) = \min_{(\mathbf{x}_i,\mathbf{x}_j)} \frac{1}{N_J}\sum_{t=1}^{N_J}(\mathbf{x}_{wi}^t(\mathbf{x}_i) - \mathbf{x}_{wj}^t(\mathbf{x}_j))^2, \tag{12}$$

with $\mathbf{x}_{wi}^t(\mathbf{x}) = \mathbf{P}_i^t(T_i^{-1}(\mathbf{x}))$, is zero, then it is assumed that there exists a joint between i and j with coordinates

$$(\mathbf{x}_i^*, \mathbf{x}_j^*) = \arg\min_{(\mathbf{x}_i,\mathbf{x}_j)} d(\mathbf{x}_i,\mathbf{x}_j). \tag{13}$$

Of course, due to noise, this value will never truly be zero. The deviation from zero can be incorporated into a confidence measure of the existence of a joint between links i and j.

As an alternative to determining the coordinates $\mathbf{x}_i$ and $\mathbf{x}_j$ one can assume that the joint centers map to the same world coordinate location at $t = 0$ and solve for

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \frac{1}{N_J}\sum_{t=1}^{N_J}(\mathbf{P}_i^t(\mathbf{x}) - \mathbf{P}_j^t(\mathbf{x}))^2. \tag{14}$$

Ideally one would use the Euclidean 3D world coordinate transforms $\mathbf{P}(\mathbf{x}) = \mathbf{R}\mathbf{x} + \mathbf{t}$ and determine joint locations in 3D (cf. [10]), however, we can only observe the projected motion of body parts in the image plane. We hence use the general 2D affine transform as obtained from the motion segmentation stage $\mathbf{P}_i^t(\mathbf{x}) = \mathbf{A}_i^t\mathbf{x} + \mathbf{t}_i^t$ wi h $\mathbf{A}_i^t \in M(2,2)$ and $\mathbf{x}, \mathbf{t}_i^t \in \mathcal{R}^2$ and obtain

$$\mathbf{x}_{ij}^* = \arg\min_{\mathbf{x}} d_{ij}(\mathbf{x}) = \arg\min_{\mathbf{x}} \frac{1}{N_J}\sum_{t=1}^{N_J}(\mathbf{A}_i^t\mathbf{x} + \mathbf{t}_i^t - \mathbf{A}_j^t\mathbf{x} - \mathbf{t}_j^t)^2. \tag{15}$$

The sum achieves its minimum at

$$\mathbf{x}_{ij}^* = -(\sum_t (\mathbf{A}_{ij}^t)^T\mathbf{A}_{ij}^t)^{-1}\sum_t (\mathbf{A}_{ij}^t)^T(\mathbf{t}_{ij}^t), \tag{16}$$

with $\mathbf{A}_{ij}^t = \mathbf{A}_i^t - \mathbf{A}_j^t$ and $\mathbf{t}_{ij}^t = \mathbf{t}_i^t - \mathbf{t}_j^t$. The *average joint coincidence* can be expressed as

$$d_{ij}^2 = \frac{1}{N_J}\sum_{t=1}^{N_J}(\mathbf{A}_{ij}^t\mathbf{x}^* - \mathbf{t}_{ij}^t)^2. \tag{17}$$

The values of $\mathbf{x}_{ij}^*$ and d_{ij} denote for pairs of possible links i and j, the location and average coincidence of a possible joint. To obtain a reliable confidence measure c_{ij} for the existence of

a joint between i and j we denote $a_{ij} = 1$ to be the event that there exists a joint between i and j and correspondingly with $a_{ij} = 0$ that there is no joint. We assume that the joint coincidence is a random variable with $p(d_{ij}|a_{ij} = 1)$ exponentially decreasing in d_{ij}

$$p(d_{ij}|a_{ij} = 1) = \frac{1}{a_d} e^{-a_d d_{ij}}. \tag{18}$$

In addition, the distance of the joint location $\mathbf{x}^*_{ij}$ from the respective segments i and j is incorporated into the confidence measure where the distance is expressed in terms of the Mahalanobis distance from the respective segment masks in the reference frame. More specifically, the distance of a joint center $\mathbf{x}$ from the i^{th} link is given as

$$s_i(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mu_i)^T \Sigma_i^{-1} (\mathbf{x} - \mu_i), \tag{19}$$

with μ_i the center of weight and Σ_i the moment matrix of the pixel mask in the reference frame that constitutes the i^{th} segment. This distance is also assumed to be a random variable that is distributed exponentially. We hence get

$$p(\mathbf{x}^*_{ij}|a_{ij} = 1) = \frac{1}{a_s^2} e^{-a_s(s_i(\mathbf{x}^*_{ij}) + s_j(\mathbf{x}^*_{ij}))}. \tag{20}$$

The parameters a_d and a_s from equations (18) and (20) should ideally be estimated from training data. For convenience we chose these values manually to be $a_d = 1.5\,\text{pixel}$ and $a_s = 1.5$. The confidence measure c_{ij} hence relies on two factors: the average coincidence of links, d_{ij} and the distance from the respective link segments in the image plane. Assuming uniform priors we can use Bayes law to obtain

$$c_{ij} = p(a_{ij} = 1|\mathbf{x}^*_{ij}, d_{ij}) \cong p(d_{ij}|a_{ij} = 1) p(\mathbf{x}^*_{ij}|a_{ij} = 1). \tag{21}$$

The problem of determining the true joints between the visible links is now to select a subset of edges of a fully connected undirected weighted graph $G = (V, E)$, where $V = \{C_i\}$ is the set of all links and E the set of edges with weights c_{ij}.

If all segments that were extracted during the motion segmentation stage would constitute links in one single model, the search for the true joints is solved by calculating the maximum spanning tree of G [10]. However, spurious segments such as the background segment can be

observed and have to be pruned from G. This problem is similar to the clustering of Moving Light Displays in [8] where a cut threshold was used to remove spurious connections in MSTs. We remove spurious links from the maximum spanning tree of the link connectivity graph G by comparing the confidences of all edges of the tree with the median of all confidences of the tree. However, if the number of observed candidate links is small, the median approach becomes unreliable and we employ simple thresholding.

5. Model Design and Tracking

From the collected information a kinematic chain model is build as follows: For the sake of simplicity we assume a two-link kinematic chain as depicted in Figure 2. A transformation needs to be defined that maps points from each link coordinate system into the image plane. A point $\hat{\mathbf{x}}_i$ is mapped from the coordinate system of the i^{th} link into the image plane through the transformation

$$\mathbf{p}(\hat{\mathbf{x}}_i, \varphi; \xi) \;=\; G_0(\varphi_0) G(\varphi_i; \xi_i) T_{ji} \hat{\mathbf{x}}_i, \tag{22}$$

where $\mathbf{T}_{ji}$ maps a point from the coordinate system of link i to the system of link j *in its initial configuration*. The transformation $\mathbf{G}(\varphi_i, \xi_i)$ then performs the joint transformation in the coordinate system of link j, where φ_i denotes the *variable* parameters of the transformation (e.g., joint angle) and ξ_i the invariant parameters (e.g., location of joint i in the link j system). Additional links lead to added terms of the form $G(\varphi_k; \xi_k) T_{lk}$ in equation (22). The transformation $\mathbf{G}_0(\varphi_0)$ is the final transformation of the kinematic chain into the image plane with parameters φ_0. For this work we allowed translation, rotation and scaling of the complete model and rotation around each joint and scaling in a single direction for each link. The scaling direction of a link was chosen as the direction that connects the parent joint (i.e., the *incoming* joint) with the center of mass of the link. Hence in Figure 2, link i can rotate around the joint that connects it with link j and scale along the indicated direction. The combined system of link i and j can rotate, translate and scale with respect to the reference (image plane) coordinate system. The allowed scaling along a given direction resembles the concept of scaled prismatic link shapes [28].

Fig. 2

Equation (22) describes the full transformation of the kinematic model into the image plane and allows to perform model based tracking. Other formulations are possible such as product of exponentials [47] in which all links reside in a global body frames. Our approach resembles more the Denavit-Hartenberg formulation [40] in that it defines relative transformations between the link frames. This approach has the advantage of being able to define local link coordinate systems that allign with the images encoding the link appearance information, allowing a fast evaluation of the matching function. The points in the i^{th} link coordinate system $\hat{\mathbf{x}}_i$ are the actual pixel coordinates of the texture image J_i that encodes the link appearance. The transformed coordinate $\mathbf{p}(\hat{\mathbf{x}}_i, \Phi; \Xi)$ on the other hand is now given as image plane pixel coordinates which allows to efficiently obtain the image matching residuals

$$\hat{r}_i(\hat{\mathbf{x}}_i, \Phi; \Xi)) = J_i(\hat{\mathbf{x}}_i) - I_t(\mathbf{p}(\hat{\mathbf{x}}_i, \Phi; \Xi)), \tag{23}$$

with $\Phi = \{\varphi_i\}$ and $\Xi = \{\xi_i\}$ the set of all variant and invariant chain parameters. To test the extracted model we implemented a particle filter that performs the model based motion capture by iteratively propagating pose hypothesis over time. We use an intuitive weighting function rather than constructing a true probabilistic likelihood function [38]. The weighting function

$$w(I_t, \varphi_t) \sim e^{-\frac{1}{L}\sum_i \frac{1}{Z_i}\sum_{\hat{\mathbf{x}}\in J_i} \hat{\lambda}_i(\hat{\mathbf{x}})(\hat{r}_i(\hat{\mathbf{x}}, I_t, \varphi_t))^2}, \tag{24}$$

with L the number of links in the model and with the normalization factor $Z_i = \sum_{\hat{\mathbf{x}}\in J_i} \hat{\lambda}_i(\hat{\mathbf{x}})$.

The $\hat{\lambda}_i(\hat{\mathbf{x}})$ denote the alpha mask information of the i^{th} link at the (link) coordinate $\hat{\mathbf{x}}$ respectively. The weighting function simply measures the matching quality of the model in a given configuration and location φ_t registered to the image I_t. To allow sub pixel accuracy, values at non-integer locations in I_t are obtained through interpolation. Particle filters are extremely good at avoiding local maxima during the tracking process, especially in situations where link displacements of magnitude comparable to the link dimensions occur, which may prove very difficult for standard registration methods based on image gradients. The tracking approach we employed performs well for moderately long image sequences. The use of scaled prismatic link transformations allows to even handle foreshortening effects to some degree but

can fail in situation where, for example, the three dimensionality of the human body causes severe changes in link shape appearance. However, the simple model employed above showed to be sufficient for the purpose of this work of which the focus is the extraction of articulated model from visual data. Many other alternative model based tracking approaches could be implemented. One may consider including additional edge or silhouette information into the tracking framework [38] or use multi-scale approaches to improve performance.

6. Experiments

We applied our method to a set of synthetic and real image sequences containing articulated motion of various complexity. As a first experiment, a synthetic walking model was generated with Poser 4.0 by Curious Labs Inc. All body parts but the left leg were removed from the model in order to eliminate occlusion artifacts. The thus obtained *walking leg* was registered with and inserted into actual video footage. This type of sequence allows to generate near-realistic sequences with the added advantage of being able to control the size, walking style and appearance of the target. Since the model walks away from the camera it undergoes substantial changes in viewpoint and scale. Fig. 3

The motion segmentation stage extracts three layers including the background, with few outliers (see Figure 3). Our procedure correctly estimates the location of the knee joint and the resulting articulated model is used to successfully track the object until it leaves the field of view after a total of 94 frames. Figure 4, shows two arm-modeling experiments. Arm modeling is important in HCI applications that require to perform hand and arm tracking for "hand as a mouse" interfaces or gesture recognition applications. For both sequence, we again restricted the problem to extract a two-link, one-joint model from the scene which requires segmentation of the sequence into three layers. In both cases, the resulting model is able to track the arm in subsequent frames. Even substantial changes in zoom and viewpoint are handled correctly by the tracker as can be seen in the Figure 4, bottom. Fig. 4 Fig. 5

The correctness of the extracted models in terms of its kinematic structure depends both on the correct detection of the joints and the precision of their locations. While the correct

detection of the joints, and hence the correctness of the kinematic topology of the extracted model, can be verified visually, the precision of the joint location estimates is hard to assess for natural sequences. We therefore generated a near realistic synthetic upper body sequence of a user moving both arms using a character animation tool [48] (see Figure 5). For this sequence the precise image coordinates of the joints of the model are available which allows to measure the accuracy of the estimated joint locations.

The model extraction algorithm was applied to the sequence with the reference frame set to frame $t = 10$. The upper body of the synthetic model does not move with respect to the background and hence the algorithm observes a five component articulated model with two components for each arm. A comparison of the estimated joint locations with the available ground truth data reveals that the precision is below five pixels for all four joints with the best location estimate (the right shoulder) having sub pixel accuracy. The time varying location errors of the joints are shown in Figure 6 and summarized in Table 1. The joint location error averaged over all frames of the test sequence and all joints is $\mathrm{Mean}(\Delta) = 3.2$ pixels with a video resolution of 640×480 pixels. The final experiment was conducted in a realistic HCI application environment where a person is standing in front of a large screen interactive display that is equipped with a set-top camera. The goal is to model the complete upper body of the user. The user is performing a short exercise of arm movements to allow the system to acquire the articulated model. Figure 7 shows a person waiving both arms while moving slightly sideways with respect to the camera. This type of motion lead to the detection and extraction of a six link articulated model containing two segments for each arm and the torso. A qualitative inspection of the extracted model and comparison with the joint locations of the synthetic model indicates that the joint locations are reasonable. The error in the joint locations are however larger than for the synthetic sequence since there is an observable asymmetric vertical placement of the shoulder joints of approximately 10 pixel. The extracted model is used successfully for tracking the arms and torso of the user through the entire image sequence.

Fig. 6

Tab. 1

7. Discussion

For the development of the proposed model acquisition and initialization method, a number of simplifying assumptions have been made that need to be addressed in future work. In particular, the number of layers, and hence the maximum number of links that can be seen by the system, is supplied by the user. Future systems need to estimate the number of layers from the data. An approach to this can be found in [49]. Furthermore, the current system extracts a *cardboard* type model in which rigid layers rotate around joints in a plane parallel to the image plane. Such a model is able to correctly model a large number of situations, especially if the amount of perspective effects (parallax, movement in z-direction) are small. For situations in which these conditions do not hold, more powerful 3D models have to be constructed or an initial simplified (e.g., cardboard) model has to be extended and adapted on-line to accommodate perspective and three dimensional effects during the tracking stage. We believe that three dimensional models can be acquired in a similar fashion without any prior structural knowledge, especially if two or more simultaneous camera views are available. Also, effects due to occlusion are not handled so far. As occlusions or uncoverings of layers takes place, the system should infer a depth ordering of the extracted links. One important next step will be to include contour information inro the model tracking framework. The link segments yield nicely defined link boundaries that can be used to easily initialize contour models for contour based tracking.

Since all parts of the target to be modeled might not be visible in the initial reference frame, the model construction has to be performed over several frames such that all parts are "seen". Limbs that do not undergo any motion or move with the background remain unmodeled. For example, consider the test sequence in Figure 4 (bottom) in which the subjects trunk does not undergo any motion relative to the background. It hence remains unmodeled by the system. Furthermore, two limbs that are connected by a joint might not move relative to each other at every frame in which case a joint extraction is infeasible and has to be delayed until relative motion occurs. The system might even decide on-line that previously assumed rigid links have to be split up because a joint has been "discovered". As an example, consider the test sequence

in Figure 4 (top) in which the arm is modeled as a rigid segment. Finally, noise and estimation errors can lead to imprecise joint locations which can lead to poor performance of the model. The presented framework does not assume that the skeletal structure of the articulated model stems from the precise minimum spanning tree of possible link connections and is able to remove spurious links. Though not attempted, this allows in general to even model multiple people simultaneously. However, since the connectivity is based on a graph without cycles, no articulated models with closed loops can be handled, which does not occur much in nature, anyway.

8. Conclusion

We have presented a method for acquiring articulated models from monocular video from the ground up by performing a combination of multi-frame motion segmentation and joint constraint detection. We have shown how the proposed system is able to determine both the kinematic structure and shape of complex articulated objects and use the obtained information to build corresponding articulated models. These models can subsequently be used for visual tracking, thus showing how the general problem of model initialization and adaptation can be solved for a wide variety of applications. Our approach can be viewed as giving the system knowledge about the building blocks (limbs and joints) of articulated motion without giving any assembly instructions. While this approach is currently not able to compete with detailed hand crafted models, it offers the potential of gaining further insight into the domain of articulated motion capture, analysis and synthesis.

Acknowledgments

This work was supported in part by the following grants: National Science Foundation CAREER Grant IIS-97-33644, NSF Grant IIS-0081935, and NSF Grant BCS-0113030.

References

[1] D. M. Gavrila, The visual analysis of human movement: A survey, *Computer Vision and Image Understanding*, vol. 73, no. 1, pp. 82–98, 1999.

[2] R. Sharma, I. Poddar, E. Ozyildiz, S. Kettebekov, H. Kim, and T. Huang, Toward intrepretation of natural speech/gesture: Spatial planning on a virtual map, in *Proceedings of the 1999 Advanced Display Federated Laboratory Symposium*, February 1999, pp. 35–39.

[3] I. Poddar, Y. Sethi, E. Ozyildiz, and R. Sharma, Toward natural gesture/speech HCI: A case study of weather narration, in *Proc. Second Workshop on Perceptual User Interface (PUI'98)*, Nov 1998, pp. 1–6.

[4] Emilio Schapira and Rajeev Sharma, Experimental evaluation of vision and speech-based multimodal interfaces, in *Workshop on Perceptive User Interfaces*. November 2001, ACM Digital Library, ISBN 1-58113-448-7.

[5] N. Vasconcelos and A. Lippman, Empirical Bayesian motion segmentation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 12, February.

[6] S. Ayer and H. Sawhney, Layered representation of motion video using robust maximum-likelihood estimation of mixture models and MDL encoding, *Proceedings of IEEE Intl. Conf. on Computer Vision*, pp. 777–784, 1995.

[7] J. Wang and E. Adelson, Representing moving images with layers, *IEEE Transactions on Image Processing*, vol. 6, no. 3, pp. 625–638, 1994.

[8] R. F. Rashid, Towards a system for the interpretation of moving light displays, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-2, no. 6, pp. 574–581, November 1980.

[9] R. J. Holt, T. S. Huang, A. N. Netravali, and R. J. Qian, Determining articulated motion from perspective views: a decomposition approach, *Pattern Recognition*, vol. 30, pp. 1435–1449, 1997.

[10] J. F. O'Brien, R. E. Bodenheimer, Jr., G. J. Brustow, and J. K. Hodkins, Automatic joint parameter estimation from magnetic motion capture data, in *Proc. Graphics Interface 2000*, 2000.

[11] Y. Song, L. Goncalves, E. Di Bernardo, and P. Perona, Monocular perception of biological motion-detection and labeling, in *Proc. Seventh IEEE International Conference on Computer Vision*, 1999, vol. 2, pp. 805–812.

[12] Y. Song, X. Feng, and P. Perona, Towards detection of human motion, in *Computer Vision and Pattern Recognition, Proceedings. IEEE Conference on*, 2000, vol. 1, pp. 810–817.

[13] Kong Man Cheung, Takeo Kanade, J.-Y. Bouguet, and M. Holler, A real time system for robust 3d voxel reconstruction of human motions, in *Proceedings of the 2000 IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head, South Carolina*, June 2000, vol. 2, pp. 714 – 720.

[14] I. A. Kakadiaris, D. Metaxas, and R. Bajcsy, Active part-decomposition, shape and motion estimation of articulated objects: A physics-based approach, in *Proc. IEEE Computer Vision and Pattern Recognition Conference, Seattle, WA*, June 1994.

[15] A. Hilton, T. Gentils, and D. Beresford, Virtual people: Capturing 3d articulated models of individual people, in *IEE Colloqium on Computer Vision for Virtual Human Modelling (Ref. No. 1998/433)*, July 1998.

[16] R. Plaenkers and P. Fua, Articulated soft objects for video-based body modeling, in *Proc. International Conference on Computer Vision, Vancouver, Canada*, 2001.

[17] Sergey Ioffe and David Forsyth, Human tracking with mixtures of trees, in *Proc. International Conference on Computer Vision, Vancouver, Canada*, 2001.

[18] S. Ju, M. Black, and Y. Yacoob, Cardboard People: a parameterized model of articulated image motion, in *Proc. International Conference on Automatic Face and Gesture Recognition*, 1996, pp. 38–44.

[19] Yaser Yacoob and Michael J. Black, Parameterized modeling and recognition of activities, *Computer Vision and Image Understanding: CVIU*, vol. 73, no. 2, pp. 232–247, 1999.

[20] L. Torres, D. Garcia, and A. Mates, A robust motion estimation and segmentation approach to represent moving images with layers., in *International Conference on Acoustics, Speech and Signal Processing*, April 1997.

[21] G. D. Borshukov, G. Bozdagi, Y. Altunbasak, and A. M. Tekalp, Motion segmentation by multistage affine classification, *IEEE Trans. Image Processing*, vol. 6, pp. 1591–1594, November 1997.

[22] A. O. Dempster, N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the em algorithm., *Journal of the Royal Statistical Society, Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.

[23] J. O'Rourke and B. U. Badler, Model-based image analysis of human motion using constraint propagation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 2, pp. 522–536, 1980.

[24] T. B. Moeslund and E. Granum, A survey of computer vision-based human motion capture, *Computer Vision and Image Understanding*, , no. 81, pp. 231–268, 2001.

[25] C. Jia-Ching and J. Moura, Tracking human walking in dynamic scenes, in *Proc. International Conference on Image Processing*, 1997, vol. 1, pp. 137–140.

[26] N. R. Howe, M. E. Leventon, and W. T. Freeman, Bayesian reconstruction of 3d human motion from a single-camera video, Tech. Rep. TR-99-37, MERL - A Mitsubishi electric research laboratory, October 1999.

[27] V. I. Pavlovic, J. Rehg, C. Tat-Jen, and K. Murphy, A dynamic Bayesian network approach to figure tracking using learned dynamic models, in *Seventh IEEE International Conference on Computer Vision*, 1999, vol. 1, pp. 94–101.

[28] Daniel D. Morris and James M. Rehg, Singularity analysis for articulated object tracking, in *Proc. IEEE Conference on Computer Vision and Pattern Recognition, Santa Barbara, CA*, June 1998.

[29] D.M. Gavrila and L.S. Davis, 3-d model-based tracking of humans in action: a multi-view approach, in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 1996, pp. 73–80.

[30] A. Wu, M.. Sha, and N. da Vitora Lobo, A virtual 3d blackboard: 3d finger tracking using a single camera, in *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 2000, pp. 536–542.

[31] T. B. Moeslund and E. Granum, Multiple cues used in model-baed human motion capture, in *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, 2000, pp. 362–367.

[32] V. Filova, F. Solina, and J. Lenarcic, Automatic reconstruction of 3d human arm motion from a monocular image sequence, *Machine Vision and Application*, vol. 10, pp. 223–231, 1998.

[33] L. Goncalves, E. D. Bernardo, E. Ursella, and P. Perona, Monocular tracking of the human arm in 3d, in *Proc. International Conference on Computer Vision*, 1995, pp. 764–770.

[34] A. Doucet, N. de Freitas, and N. Gordon, *Sequential Monte Carlo Methods in Practice*, Statistics for Engineering and Information Science. Springer Verlag, New York, Heidelberg, Berlin, 2001.

[35] Hedvig Sidenbladh, Michael J. Black, and David J. Fleet, Stochastic tracking of 3d human figures using 2d image motion, in *In European Conference on Computer Vision, Dublin, Ireland*, November 2000, vol. 2, pp. 702–718.

[36] M. Isard and A. Blake, Condensation - conditional density propagation for visual tracking, *International Journal of Computer Vision*, vol. 1, no. 29, pp. 5–28, 1998.

[37] Hedvig Sidenbladh and Michael J. Black, Learning image statistics for bayesian tracking, in *Proc. IEEE International Conference on Computer Vision, Vancouver*, July 2001, vol. 2, pp. 709–716.

[38] J. Deutscher, A. Blake, and I. Reid, Articulated body motion capture by annealed particle filtering, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 2000, vol. 2, pp. 126–133.

[39] C. Bregler and J. Malik, Tracking people with twists and exponential maps, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 1998.

[40] R. M. Murray, Z. Li, and S. S. Sastry, *A Mathematical Introduction to Robotic Manipulation*, CRC Press, first edition, 1994.

[41] M. M. Covell, Ali Rahimi, M. Harville, and T. J. Darrell, Articulated-pose estimation using brightness and depth-constancy constraints, in *Proc. Conference on Computer Vision and Pattern Recognition*, June 2000.

[42] A. Ude and M. Riley, Prediction of body configuration and appearance for model-based estimation of articulated human motions, in *Proc. IEEE Int. Conf. Systems, Man, and Cybernetics*, 1999, pp. 687–691.

[43] T. Drummond and R. Cipolla, Real-time tracking of highly articulated structures in the presence of noisy measurements, in *Proc. International Conference on Computer Vision, Vancouver, Canada*, 2001.

[44] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, first edition, 1999.

[45] B. D. Lucas and T. Kanade, An iterative image registration technique with an application to stereo vision, in *Proc. Interrnational Joint Conference on Artificial Intelligence*, 1981.

[46] Jianbo Shi and Carlo Tomasi, Good features to track, in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'94), Seattle, Washington*, June 1994.

[47] C. Bregler and J. Malik, Tracking people with twists and exponential maps, in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 1998, pp. 8–15.

[48] Curious Labs, Inc., 655 Capitola Road, Suite 200, Santa Cruz, CA 95062, USA, http://www.curiouslabs.com, *Poser 4.0*.

[49] P. H. S. Torr, R. Szeliski, and P. Anandan, An integrated Bayesian approach to layer extraction from image sequences, March 2001, vol. 23, pp. 297–303.

Figures

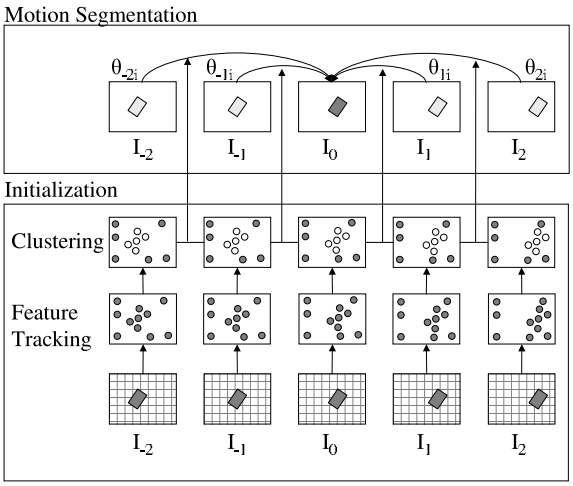

Figure 1: Summary of the initialization and motion segmentation procedure. The motion segmentation that is performed after an initial feature tracking and track clustering stage for bootstrapping leads to segmentation information for a central frame and motion information for each obtained segment and each frame used for the motion segmentation.

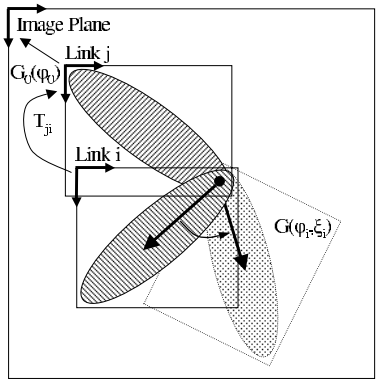

Figure 2: Definition of the coordinate systems and transformation for the extracted kinematic chain.

Figure 3: Sequence showing the leg portion of a synthetic walker registered into the world coordinates of actual video footage. The top right image shows the extracted layers with. The bottom two images show two frames from the resulting tracking sequence with the extracted articulated model.

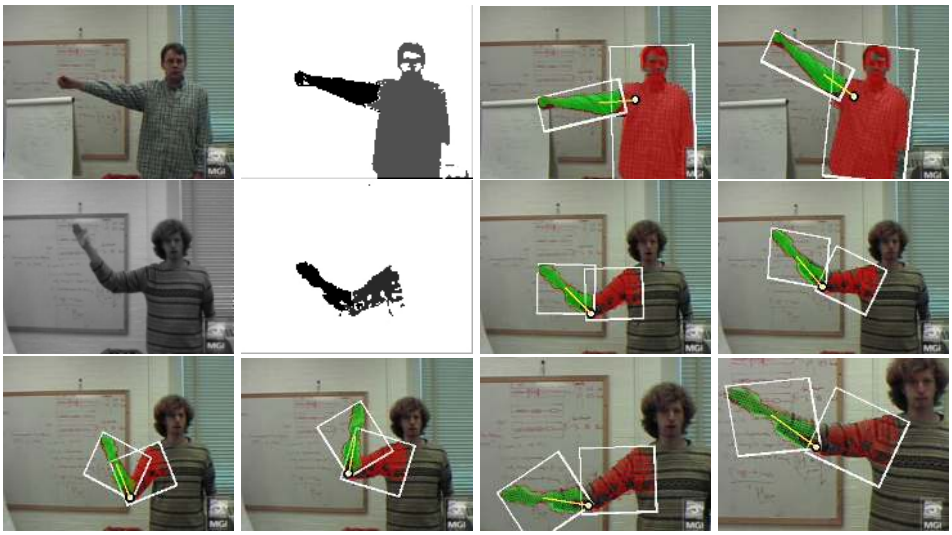

Figure 4: Arm model acquisition and tracking. Top row: Subject moving with respect to the camera exercising the shoulder joint. Bottom rows: Subject exercising elbow and shoulder joint while not moving with respect to the camera. The latter example shows subsequent tracking while the camera zooms and changes viewing direction.

Figure 5: Synthetic image sequence of a person moving both arms. The sequence is modelled with five motion segments shown in the top right image. White pixels denote outliers. The joint locations that were estimated for this sequence are compared to the true locations in Table 1.

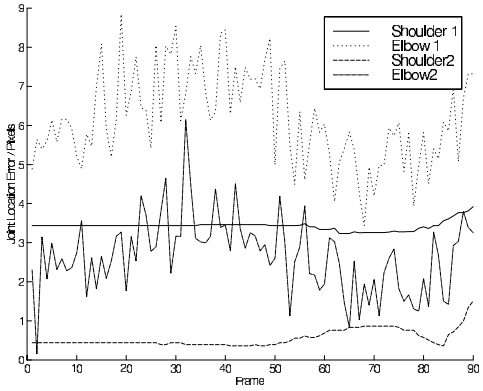

Figure 6: Precision of joint location estimates for the synthetic image sequence shown in Figure 5.

	Left Arm		Right Arm	
	Shoulder	Elbow	Shoulder	Elbow
Δ_x	-0.97	-4.76	0.28	2.23
Δ_y	3.30	-1.08	0.35	-0.58
$\Delta(t=0)$	3.44	4.88	0.44	2.30
$\mathrm{Mean}(\Delta)$	3.43	6.17	0.56	2.67

Table 1: Precision of joint location estimates for the sequence shown in Fig. 5 compared to ground truth data. The values Δ_x and Δ_y denote the deviation in x and y-direction respectively, while Δ denotes the L_2 norm of (Δ_x, Δ_y). All values are in units of pixels in a video of resolution 640×480.

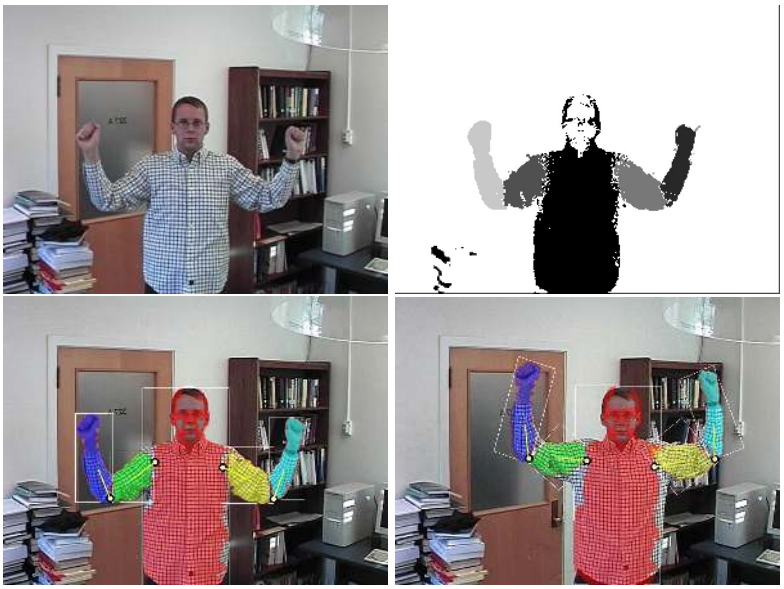

Figure 7: Subject with waiving arms and swaying torso.