Machine Vision and Applications (2003) 14: 218–228
Digital Object Identifier (DOI) 10.1007/s00138-002-0081-2

Machine Vision and Applications

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, USA
e-mail: {krahnsto,yeasin,rsharma}@cse.psu.edu

Published online: 8 August 2003 –

Abstract. Tracking, classification and visual analysis of articulated motion is challenging because of the difficulties involved in separating noise and variabilities caused by appearance, size and viewpoint 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. Image sequences are decomposed into rigid components that can undergo parametric motion. The relative motion of 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.

Key words: Model assembly – Model-based visual tracking – Joint detection – Model acquisition – Articulated motion

1 Introduction

Important capabilities of vision-based human–computer interaction (HCI) 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 viewpoint dependent variabilities of the visual motion patterns. The use of explicit articulated models is promising for overcoming these challenges because it allows directly encoding 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 (Gavrila 1999). 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 handcrafted. 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 model is registered to the target in the first frame. This problem of model initialization is commonly reported to be performed manually by the user. In summary, the use of model-based motion capture systems in many domains (e.g., surveillance, HCI, 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 HCI and gesture recognition applications (Kettebekov et al. 2000; Poddar et al. 1998; Schapira and Sharma 2001) 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 single-view video, assuming only the concept of articulated motion as prior knowledge. We assume that the input data can be viewed as a collection of rigid elements 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 (Vasconcelos and Lippman 2001; Ayer and Sawhney 1995; Wang and Adelson 1994)

Correspondence to: N. Krahnstoever

to decompose simultaneously 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 constitute 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 by using 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 Sect. 2. Following this, we present our approach to extract the rigid components of the input sequence that form the link candidates in Sect. 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 Sect. 5. Experiments on real and synthetic data are presented in Sect. 6, followed by a discussion of these results in Sect. 7. Finally, Sect. 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), and passive or active markers, or be extracted from image sequences using feature trackers.

Early work by Rashid (1980) presents a comprehensive algorithm for analyzing MLDs. Point features were clustered into objects using a minimum spanning tree approach together with a cut criterion for splitting the resulting tree into clusters. The underlying skeletal structure of the point features was obtained by calculating the minimum spanning tree on each group. Rashid stressed the importance of velocity information in obtaining robust estimates of skeletal structure.

Using magnetic motion capture data obtained with magnetic sensors, O'Brien et al. (2000) 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 minimum spanning tree 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. Barrón and Kakadiaris (2001) present a semi-automatic approach for extracting human skeletal models from single images. Their method requires the manual selection of a set of landmarks that are used in combination with human body shape statistics and joint limit constraints to infer the anthropometry and pose of the target.

A motion modeling approach not based on the articulated structure of the human body was developed by Song et al. (1999, 2000). 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 and used for subsequent motion analysis.

2.2 Human-shape acquisition

The acquisition of precise human body shape models has so far mostly been investigated for situations in which multiple camera views are available. Cheung et al. (2000) acquired human body models without any prior structural assumptions using a large number of camera views in a customized laboratory environment. Kakadiaris and Metaxas (1998) and Hilton et al. (1998) presented acquisition systems that require the views of at least three cameras to perform the modeling process. The latter work uses a generic body model onto which the image of a person's view is mapped. The work presented in (Plaenkers and Fua 2001) 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 video and low-cost HCI applications. Furthermore, a flexible acquisition procedure should not be restricted to an a priori given articulated structure.

Ioffe and Forsyth (2001) 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 acquiring 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 (Ju et al. 1996; Yacoob and Black 1999).

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 (Ayer and Sawhney 1995; Vasconcelos and Lippman 2001; Torres et al. 1997; Borshukov et al. 1997; Wang and Adelson 1994). 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 (1994) and subsequent improvements (Torres et al. 1997; Borshukov et al. 1997) approached the problem using clustering, the problem has since been formulated in a Bayesian framework using an expectation–maximization (EM) algorithm (Dempster et al. 1977).

2.4 Model-based tracking

The articulated models acquired by our proposed method are to be used in the context of model-based motion tracking in HCI applications. The approach of model-based human tracking has been pioneered by O'Rourke and Badler (1980). 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 (Moeslund and Granum 2001).

Ju et al. (1996) approximate humans with cardboard models, which are basically 2D models specialized for 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 (Jia-Ching and Moura 1997) and handle occlusion (Howe et al. 1999).

Pavlovic et al. (1999) address the problem of learning dynamics from training data in a Bayesian framework. They also employ a 2D model, but use scaled prismatics (Morris and Rehg 1998) that are able to handle 3D foreshortening effects and avoid singularities common in 3D kinematic modeling approaches.

The resulting models that we acquire from video are also of the scaled prismatic cardboard type, but future work is aimed at improving the proposed approach to obtain full 3D models.

Full-body 3D model tracking has been demonstrated by Gavrila and Davis (1996), who developed a four-camera tracking system using a model consisting 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.

Bregler and Malik (1998) developed an articulated motion capture framework that parameterizes the kinematic chain of the human body in an exponential twist formulation pioneered in robotics (Murray et al. 1994). The authors take a variational approach that relies on matching the appearance of limbs with the image content. The formulation is ultimately based on the brightness constancy assumption and linear approximations of the twists. They were able to report good tracking results on monocular and multi-view image sequences. Other research that utilizes the twist-formulation was presented by Ude and Riley (1999) and Covell et al. (2000).

HCI applications often only require the modeling of the upper body, or one or both arms (Schapira and Sharma 2001). Model-based arm tracking is particularly promising in HCI application and has been addressed in (Wu et al. 2000; Moeslund and Granum 2000; Filova et al. 1998; Goncalves et al. 1995). In all cases, two-link models were used.

To verify the utility of the articulated models acquired by our proposed work, we implemented a simple sequential Monte Carlo filter (Doucet et al. 2001; Isard and Blake 1998) as described in Sect. 5. Sequential Monte Carlo methods are gaining popularity due to their simplicity and robustness towards noise and clutter.

Using a particle filtering approach, Sidenbladh et al. (2000) perform 3D reconstruction of human motion observed with a single camera using models consisting of ten cylinders under perspective projection connected by joints. Appearance information of the cylinders is adapted incrementally from the image sequences. The authors encouraged the use of more persistent appearance models. Sidenbladh and Black (2001) also 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 works show the benefit of finely tuned models that are learned or acquired from data and motivate the work proposed in this paper.

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 EM algorithm (Dempster et al. 1977).

In this approach, motion segmentation is performed iteratively in two stages until convergence: expectation and maximization. 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. The extraction of good initial values is performed during the initialization stage and discussed in detail in Sect. 3.2.

Our approach to performing the motion segmentation is an extension of the works of Vasconcelos and Lippman (2001), and Ayer and Sawhney (1995), which are both two-frame algorithms. 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 sufficient motion that distinguishes itself from the motion of all other segments as far as motion parameters are concerned. If two limbs perform the same or very similar motion in a considered time interval, they cannot be distinguished from each other and will be viewed as one part. Therefore, a two-frame approach provides a too small observation frame for the video data considered in this work. In order to increase the chance of observing distinguishable motion patterns for

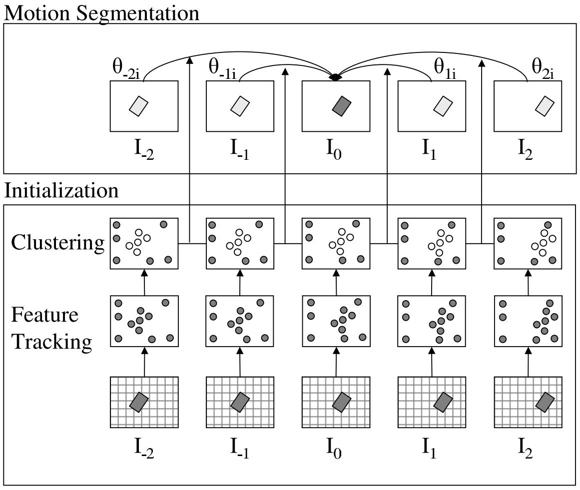

Fig. 1. Summary of the initialization and motion segmentation procedure. The motion segmentation that is performed after an initial feature-tracking and track-clustering stage for bootstrapping leads to segmentation information for a central frame and motion information for each obtained segment and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

pairs of segments, our algorithm therefore performs the motion segmentation on one frame, like existing approaches, but draws 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 Fig. 1): The motion segmentation is performed for a reference frame I_0 based on a set of N_F previous and subsequent frames $\mathbf{I}_{N_F} = \mathbf{I}^+_{N_F} \cup \mathbf{I}^-_{N_F}$ with $\mathbf{I}^\pm_{N_F} = \{I_{\pm 1}, \ldots, I_{\pm N_F}\}$. To obtain good segmentations, the value of N_F has to be choosen such that sufficient segment motion occurs during the considered image sequence. Values between 2 and 6 were found to be sufficient for the considered sequences. The algorithm estimates motion parameters θ_{if} that map the ith 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 layer L_i. In order to handle effects caused by occlusion, the motion segmentation is performed separately in forward and backward directions and the results are combined in a final stage. Only the forward case is outlined below.

Using Bayes rule we have (cf. Vasconcelos and Lippman (2001)):

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

with $\mathbf{\Theta}_i = \{\theta_{if}, f = 1, \ldots, N_F\}$ and Ψ_i denoting additional shape and color parameters to be specified shortly and c a normalization factor. 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 the proposed motion segmentation implementation, this term draws its information from three sources:

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

In Eq. (2), $P_{\mathrm{r}}(.)$ models the residuals arising from the match between I_0 and the following images given the motion parameters, and $P_{\mathrm{s}}(.)$ and $P_{\mathrm{c}}(.)$ express the conformance to the shape and color model of the ith layer respectively. The residual term $P_{\mathrm{r}}(.)$ 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 $\mathbf{I}^+_{N_F}$,

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

with residuals

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

The function $\mathbf{F}(\mathbf{x}; \theta)$ 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 Eq. (2) assumes that the pixels in each layer are normally distributed around a center location μ_i with empirical shape covariance matrix Σ_i and is written as

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

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 textureless 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 between pixels in a given layer. We assume that pixel values are normally distributed in RGB color space according to the following relationship:

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

This term helps to improve the assignment of pixels to layers, especially at the boundaries of layer regions and in textureless 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 2D affine transform is used

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

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 $\mathbf{x}$ (Ayer and Sawhney 1995) or used to impose smoothness on the layers (Vasconcelos and Lippman 2001). We follow the latter option and use a MRF prior (Vasconcelos and Lippman 2001) that effectively enforces a spatial coherence in the layer assignments and leads to smoother results.

At each E-step, the calculation of Eq. (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. Note that Eq. (10) separates into a sum of independent terms. Equation (10) is minimized using Gauss–Newton optimization with line search (Nocedal and Wright 1999), with analytically calculated gradients and Hessians. The EM iteration has to be repeated until convergence.

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 (Lucas and Kanade 1981; Shi and Tomasi 1994) 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 only in the direction of the layer movement. It can hence be canceled out by performing motion segmentation in both forward and backward directions 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 region and the image content is 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 ith link coordinate system $\mathbf{x}_i = \mathbf{T}_i(\mathbf{x}_w)$ at time $t = 0$. With $\mathbf{F}_i^f$ 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 ith link, $\mathbf{x}^t_{wi} = \mathbf{F}^t_i(\mathbf{x}_w)$.

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, $\mathbf{F}^t_i(\mathbf{T_i}^{-1}(\mathbf{x_i})) = \mathbf{F}^t_j(\mathbf{T_j}^{-1}(\mathbf{x}_j))$, for all t. The points $\mathbf{x_i}$ and $\mathbf{x_j}$ 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 move non-uniformly with respect to each other over extended periods of time, it is reasonable to assume that this indicates the existence of a joint. More specifically, if the average *link coincidence*

$$
\begin{aligned}
d^2_{ij} &= \min_{(\mathbf{x}_i,\mathbf{x}_j)} d(\mathbf{x}_i,\mathbf{x}_j) \\
&= \min_{(\mathbf{x}_i,\mathbf{x}_j)} \frac{1}{N_J}\sum_{t=1}^{N_J}(\mathbf{x}^t_{wi}(\mathbf{x}_i) - \mathbf{x}^t_{wj}(\mathbf{x}_j))^2, \qquad (12)
\end{aligned}
$$

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

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

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

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \frac{1}{N_J}\sum_{t=1}^{N_J}(\mathbf{F}^t_i(\mathbf{x}) - \mathbf{F}^t_j(\mathbf{x}))^2. \qquad (14)$$

Ideally one would use the Euclidean 3D world coordinate transforms $\mathbf{F}(\mathbf{x}) = \mathbf{R}\mathbf{x} + \mathbf{t}$ and determine joint locations in 3D (cf. O'Brien et al. (2000)); however, we can only observe the projected motion of body parts in the image plane. Hence we use the general 2D affine transform as obtained from the motion segmentation stage $\mathbf{F}^t_i(\mathbf{x}) = \mathbf{A}^t_i\mathbf{x}+\mathbf{t}^t_i$ with $\mathbf{A}^t_i$ a 2×2 matrix and $\mathbf{x}, \mathbf{t}^t_i \in \mathcal{R}^2$ and obtain

$$
\begin{aligned}
\mathbf{x}^*_{ij} &= \arg\min_{\mathbf{x}} d_{ij}(\mathbf{x}) \\
&= \arg\min_{\mathbf{x}} \frac{1}{N_J}\sum_{t=1}^{N_J}(\mathbf{A}^t_i\mathbf{x} + \mathbf{t}^t_i - \mathbf{A}^t_j\mathbf{x} - \mathbf{t}^t_j)^2. \qquad (15)
\end{aligned}
$$

The sum achieves its minimum at

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

with $\mathbf{A}^t_{ij} = \mathbf{A}^t_i - \mathbf{A}^t_j$ and $\mathbf{t}^t_{ij} = \mathbf{t}^t_i - \mathbf{t}^t_j$. The average *joint coincidence* can be expressed as

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

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 $a_{ij} = 0$ we denote 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}

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

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 ith link is given as

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

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

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

The parameters a_d and a_s from Eqs. (18) and (20) should ideally be estimated from training data. For convenience we chose these values manually to be $a_d^{-1} = 1.5$ 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

$$
\begin{aligned}
c_{ij} &= p(a_{ij} = 1|\mathbf{x}^*_{ij}, d_{ij}) \\
&\propto p(d_{ij}|a_{ij} = 1)p(\mathbf{x}^*_{ij}|a_{ij} = 1). \qquad (21)
\end{aligned}
$$

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 is 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 (O'Brien et al. 2000). 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 MLDs in (Rashid 1980) where a cut threshold was used to remove spurious connections in maximum spanning trees. 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 assembly and tracking

From the collected information a kinematic chain model is built as follows: For the sake of simplicity we assume a two-link kinematic chain as depicted in Fig. 2. A transformation

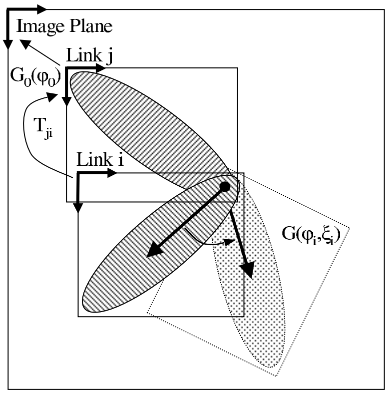

Fig. 2. Definition of the coordinate systems and transformation for the extracted kinematic chain

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 ith link into the image plane through the transformation

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

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 Eq. (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, 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 Fig. 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 (Morris and Rehg 1998).

Equation (22) describes the full transformation of the kinematic model into the image plane, allowing model-based tracking. Other formulations are possible, such as product of exponentials (Bregler and Malik 1998) in which all links reside in a global body frames. Our approach resembles more the Denavit–Hartenberg formulation (Murray et al. 1994) 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 align with the images encoding the link appearance information, allowing a fast evaluation of the matching function. The points in the ith 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 $\mathbf{f}(\hat{\mathbf{x}}_i, \Phi; \Xi)$ on the other hand is now given as image plane pixel coordinates which allows efficiently obtaining the image-matching residuals

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

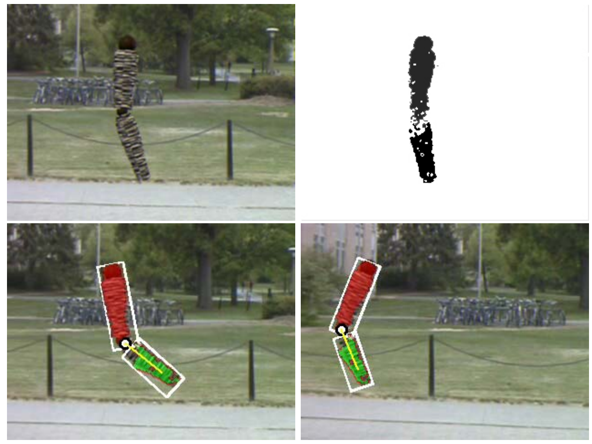

Fig. 3. Sequence showing the leg portion of a synthetic walker registered into the world coordinates of actual video footage. The *top right image* shows the extracted layers with *black pixels* denoting outliers. The *second row* shows two sample frames from the resulting tracking sequence with the extracted articulated model. The detected joint is indicated by the *white circle*

with $\Phi = \{\varphi_i\}$ and $\Xi = \{\xi_i\}$, the set of all variant and invariant chain parameters respectively. The number of estimated motion parameters depends on the structure of the model (i.e., number of joints and links) and includes global location, pose and scale of the model, in addition to one joint parameter and prismatic link scale parameter for each joint.

To test the extracted model, we implemented a particle filter that performs the model-based motion capture by iteratively propagating pose hypothesis over time. The likelihood function is based on the matching residuals

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

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 $\hat{\lambda}_i(\hat{\mathbf{x}})$ denotes the alpha mask information of the ith link at the (link) coordinate $\hat{\mathbf{x}}$. The weighting function simply measures the registration error 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 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 evenly handles foreshortening effects to some degree but can fail in situations where, for example, the three dimensionality of the human body causes severe changes in link shape appearance. However, the simple model employed above proved 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 (Deutscher et al. 2000), or the use of multi-scale approaches to improve performance.

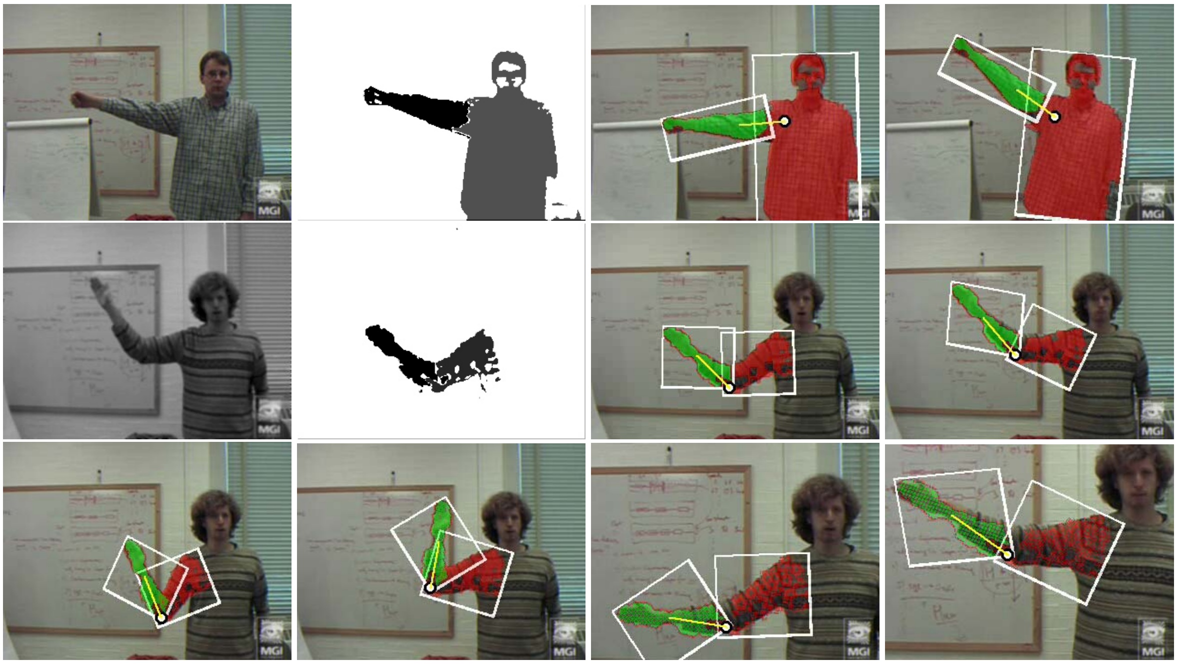

Fig. 4. Arm model acquisition and tracking. *Top row*: Subject moving with respect to the camera exercising the shoulder joint. *Bottom two rows*: Subject exercising elbow and shoulder joint while not moving with respect to the camera. The latter example shows subsequent tracking while the camera zooms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

6 Experiments

We applied our method to a set of synthetic and real image sequences containing articulated motion of various degrees of complexity. In the 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 actual video footage. This type of sequence allows the generation of 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.

The motion segmentation stage extracts three layers including the background, with few outliers (see Fig. 3). The proposed procedure correctly estimates the location of the knee joint and the resulting assembled 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 performing hand and arm tracking for "hand as mouse" interfaces or gesture recognition applications. For both sequences, we again extracted two-link, one-joint models from the scenes, which requires segmentation of the sequences 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 bottom row of Fig. 4.

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 synthetic upper body sequence of a user moving both arms using a character animation tool (Labs 2001) (see Fig. 5). For this sequence the precise image coordinates of the joints of the model are available, allowing the measurement of 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 nine 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 Fig. 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 difference in accuracy between the shoulder and elbow angles stems from fact that the shoulders are attached to the large body segment that due to its size registers much better over time than the smaller arm segments. The difference in accuracy between the right and the left arm is coincidential and stems from the randomness induced by the feature tracker and clustering during the initialization stage.

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

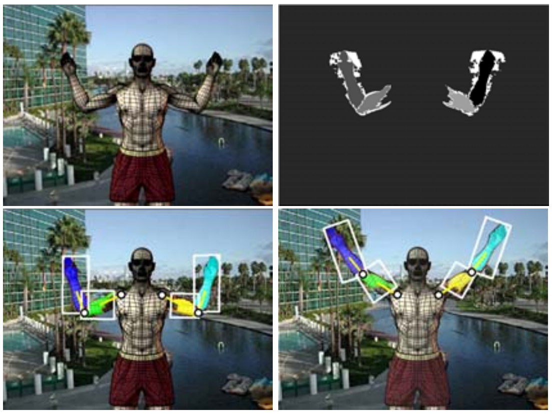

Fig. 5. Synthetic image sequence of a person moving both arms. The sequence is modelled with five motion segments shown in the *top right image*. *White pixels* denote outliers. The joint locations that were estimated for this sequence are compared to the true locations in Table 1

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 with respect to a video resolution of 640×480

	Left Arm		Right Arm	
	Shoulder	Elbow	Shoulder	Elbow
Δ_x	-0.97 pels -3.72 mm	-4.76 pels -18.3,mm	0.28 pels 1.08 mm	2.23 pels 8.56 mm
Δ_y	3.30 pels 12.7 mm	-1.08 pels -4.15 mm	0.35 pels 1.34 mm	-0.58 pels 8.56 mm
$\Delta_t\|_{t=0}$	3.44 pels 13.2 mm	4.88 pels 18.7 mm	0.44 pels 1.69 mm	2.30 pels 8.83 mm
Mean	3.43 pels 13.2 mm	6.17 pels 23.7 mm	0.56 pels 2.15 mm	2.67 pels 10.3 mm

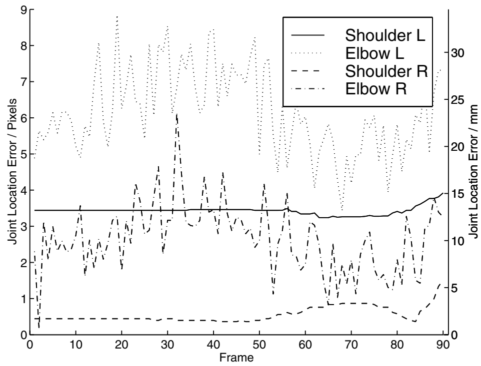

Fig. 6. Precision of joint location estimates for the synthetic image sequence shown in Fig. 5

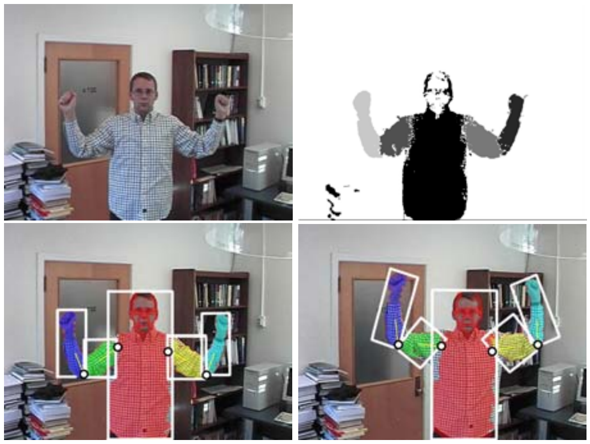

Fig. 7. Subject with waving arms and swaying torso

to allow the system to acquire the articulated model. Figure 7 shows a person waving 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 pixels. The extracted model is used successfully for tracking the arms and torso of the user through the entire image sequence.

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 (Torr et al. 2001). 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 online to accommodate perspective and three-dimensional effects during the tracking stage. We believe that three-dimensional models can be acquired from video in a similar fashion without any prior structural knowledge. 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 into 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 Fig. 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 online that previously assumed rigid links have to be split up because a joint has been "discovered". As an example, consider the test sequence in Fig. 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

This paper presents 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. It was shown that 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 were subsequently used for visual tracking, thus showing how the general problem of model initialization and adaptation can be solved for a wide variety of applications. The proposed 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 handcrafted models, it offers the potential of gaining further insight into the domain of articulated motion capture, analysis and synthesis.

Acknowledgements. 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

Ayer S, Sawhney H (1995) Layered representation of motion video using robust maximum-likelihood estimation of mixture models and MDL encoding. In: Proceedings of the Fifth International Conference on Computer Vision, Boston, June 20–23, IEEE Comput. Soc., Los Alamitos, CA, USA, pp 777–784

Barrón C, Kakadiaris IA (2001) Estimating anthropometry and pose from a single image. Comput Vision Image Understanding 81: 269–284

Borshukov GD, Bozdagi G, Altunbasak Y, Tekalp AM Motion segmentation by multistage affine classification. IEEE Trans Image Process 6:1591–1594

Bregler C, Malik J. (1998) Tracking people with twists and exponential maps. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Santa Barbara, CA, USA, pp 8–15

Cheung KM, Kanade T, Bouguet JY, Holler M (2000) 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, Vol 2, pp 714–720

Covell MM, Rahimi A, Harville M, Darrell TJ (2000) Articulated-pose estimation using brightness and depth-constancy constraints. In: Proceedings of the 2000 IEEE Conference on Computer Vision and Pattern Recognition, Vol 2, Hilton Head Island, SC, USA pp 438–445

Dempster AO, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J Royal Statistical Soc, Series B (Methodological), 39(1):1–38

Deutscher J, Blake A, Reid I (2000) Articulated body motion capture by annealed particle filtering. In: Proceedings of the 2000 IEEE Conference on Computer Vision and Pattern Recognition, Vol 2, Hilton Head Island, SC, USA, pp 126–133

Doucet A, de Freitas N, Gordon N (2001) Sequential Monte Carlo methods in practice. Springer, Berlin Heidelberg New York

Filova V, Solina F, Lenarcic J (1998) Automatic reconstruction of 3D human arm motion from a monocular image sequence. Mach Vision Appl 10:223–231

Gavrila DM (1999) The visual analysis of human movement: a survey. Comput Vision Image Understanding 73(1):82–98

Gavrila DM, Davis LS (1996) 3D model-based tracking of humans in action: a multi-view approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco pp 73–80

Goncalves L, Bernardo ED, Ursella E, Perona P (1995) Monocular tracking of the human arm in 3D. In: Proceedings of the Fifth International Conference on Computer Vision, Cambridge, MA, USA, pp 764–770

Hilton A, Gentils T, Beresford D (1998) Virtual people: capturing 3D articulated models of individual people. In: IEE Colloquium on Computer Vision for Virtual Human Modelling (Ref. No. 1998/433), July, IEE, London, UK

Howe NR, Leventon ME, Freeman WT (1999) Bayesian reconstruction of 3D human motion from a single-camera video. Technical report TR-99-37, Mitsubishi electric research laboratories, October

Ioffe S, Forsyth D (2001) Human tracking with mixtures of trees. In: Proceedings of the International Conference on Computer Vision, Vancouver, BC, IEEE Comput. Soc., Los Alamitos, CA, USA, pp 690-695

Isard M, Blake A (1998) CONDENSATION – conditional density propagation for visual tracking. Int J Comput Vision 1(29):5–28

Jia-Ching C, Moura J (1997) Tracking human walking in dynamic scenes. In: Proceedings of the International Conference on Image Processing Vol 1, IEEE, Santa Barbara, CA, USA, pp 137–140

Ju S, Black M, Yacoob Y (1996) Cardboard people: a parameterized model of articulated image motion. In: Proceedings of the International Conference on Automatic Face and Gesture Recognition, Killington, VT, USA, pp 38–44

Kakadiaris IA, Metaxas D (1998) 3D human body model acquisition from multiple views. Int J Comput Vision 30(3):191–218

Kettebekov S, Krahnstoever N, Leas M, Polat E, Raju H, Schapira E, Sharma R (2000) i^2map: crisis management using a multimodal interface. In: ARL Federate Laboratory 4th Annual Symposium, College Park, MD, March.

Curious Labs (2001) Poser 4.0. 655 Capitol Road, Suite 200, Santa Cruz, CA 95062, USA, http://www.curiouslabs.com

Lucas BD, Kanade T (1981) An iterative image registration technique with an application to stereo vision. In: Proceedings of the 7th International Joint Conference on Artificial Intelligence, Vancouver, Canada, Morgan Kaufmann, San Francisco, USA, pp 674–679

Moeslund TB, Granum E (2000) Multiple cues used in model-based human motion capture. In: Fourth IEEE International Conference on Automatic Face and Gesture Recognition, Grenoble, France, pp 362–367

Moeslund TB, Granum E (2001) A survey of computer vision-based human motion capture. Comput Vision Image Understanding 81: 231–268

Morris DD, Rehg JM (1998) Singularity analysis for articulated object tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Santa Barbara, Calif., pp 289–296

Murray RM, Li Z, Sastry SS (1994) A mathematical introduction to robotic manipulation, 1st edn. CRC Press, Boca Raton, Florida

Nocedal J, Wright SJ (1999) Numerical optimization, 1st edn. Springer, Berlin Heidelberg New York

O'Brien JF, Bodenheimer RE Jr, Brostow GJ, Hodkins JK (2000) Automatic joint parameter estimation from magnetic motion capture data. In: Proceedings of graphics Interface 2000, Montreal, Quebec, May 15–17, Peters, Natick, MA, USA, pp 53–60

O'Rourke J, Badler BU (1980) Model-based image analysis of human motion using constraint propagation. IEEE Trans Pattern Anal Mach Intell 6(2):522–536

Pavlovic VI, Rehg J, Tat-Jen C, Murphy K (1999) A dynamic Bayesian network approach to figure tracking using learned dynamic models. In: Proceedings of the International Conference on Computer Vision, Vol 1, IEEE, Kerkyra, Greece, pp 94–101

Plaenkers R, Fua P (2001) Articulated soft objects for video-based body modeling. In: Proceedings of the International Conference on Computer Vision, Vancouver, BC, IEEE Comput. Soc., Los Alamitos, CA, USA, pp 394–401

Poddar I, Sethi Y, Ozyildiz E, Sharma R (1998) Toward natural gesture/speech HCI: a case study of weather narration. In: Proceedings of the Second Workshop on Perceptual User Interface (PUI'98), San Francisco, CA, USA, pp 1–6

Rashid RF (1980) Towards a system for the interpretation of moving light displays. IEEE Trans on Pattern Anal Mach Intell 2(6):574–581

Schapira E, Sharma R (2001) Experimental evaluation of vision and speech-based multimodal interfaces. In: Workshop on Perceptive User Interfaces, ISBN 1-58113-448-7, ACM Digital Library

Shi J, Tomasi C (1994) Good features to track. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, pp 593–600

Sidenbladh H, Black MJ (2001) Learning image statistics for bayesian tracking. In: Proceedings of the International Conference on Computer Vision, Vancouver, BC, pp 709–716

Sidenbladh H, Black MJ, Fleet DJ (2000) Stochastic tracking of 3D human figures using 2D image motion. In: European Conference on Computer Vision, Dublin, Springer Berlin Heidelberg New York, pp 702–718

Song Y, Feng X, Perona P (2000) Towards detection of human motion. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hilton Head Island, SC, USA, pp 810–817

Song Y, Goncalves L, Di Bernardo E, Perona P. (1999) Monocular perception of biological motion-detection and labeling. In: Proceedings of the International Conference on Computer Vision, IEEE, Fort Collins, CO, USA, pp 805–812

Torr PHS, Szeliski R, Anandan P (2001) An integrated Bayesian approach to layer extraction from image sequences. IEEE Trans Pattern Anal Mach Intell 23(3):297–303

Torres L, Garcia D, Mates A (1997) A robust motion estimation and segmentation approach to represent moving images with layers. In: International Conference on Acoustics, Speech and Signal Processing, Munich, 21–24 April, Vol 4, IEEE Comput. Soc., Los Alamitos, USA, pp 2981–2984

Ude A, Riley M (1999) Prediction of body configuration and appearance for model-based estimation of articulated human motions. In: Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics, Tokyo, Japan, pp 687–691

Vasconcelos N, Lippman A (2001) Empirical Bayesian motion segmentation. IEEE Trans Pattern Anal Mach Intell 23(12):217–221

Wang J, Adelson E (1994) Representing moving images with layers. IEEE Trans Image Process 6(3):625–638

Wu A, Sha M, da Vitora Lobo N (2000) A virtual 3D blackboard: 3D finger tracking using a single camera. In: Fourth IEEE International Conference on Automatic Face and Gesture Recognition, Grenoble, France, pp 536–542

Yacoob Y and Black MJ (1999) Parameterized modeling and recognition of activities. Comput Vision Image Understanding 73(2): 232–247

Nils Krahnstoever is currently pursuing a PhD degree at the Department of Computer Science and Engineering at the Pennsylvania State University. He obtained his MS in Physics at the Christian-Albrechts-University of Kiel, Germany in 1998. He has worked for the Philips Research Laboratories in Hamburg, Germany and for Advanced Interface Technologies, Inc. His research interests include human–computer interfaces based on computer vision, in particular the detection, analysis and classification of articulated motion and visual tracking.

Mohammed Yeasin joined the faculty of the Pennsylvania State University, University Park, in July 2000. Before joining Pennsylvania State University he served in the Electro-technical laboratory, Japan, as a center of excellence (COE) research fellow. He obtained his PhD in degree in electrical engineering from the Indian Institute of Technology, Bombay, in May 1998. Prior to this he also served in the Electrical and Electronic Engineering department, Bangladesh Institute of Technology (BIT), Chittagong, as a lecturer. Dr. Yeasin's chief interest is in the role of computational vision in achieving robust behavior by autonomous systems. The goal is to address issues that arise in integrating computer vision with systems that benefit from real-time feedback. Such systems include robots operating in a dynamic and uncertain environment, bio-medical systems and advanced human–computer interface (HCI) systems.

Rajeev Sharma is an associate professor of computer science and engineering at Pennsylvania State University. Dr. Sharma received a PhD from University of Maryland, College Park, in 1993. Prior to joining Pennsylvania State University, Dr. Sharma served as a Beckman Fellow at the University of Illinois at Urbana-Champaign for three years. Dr. Sharma's main research focus has been on the use of computer vision to develop novel human–computer interaction and intelligence-gathering techniques. Dr. Sharma's research at is been funded through several grants from National Science Foundation and Department of Defense. Dr. Sharma is also helping in transferring the results of the university research to commercial applications through a State College, Pennsylvania-based company, Advanced Interfaces (www.AdvancedInterfaces.com) in partnership with Pennsylvania State University. Advanced Interfaces is a leader in new intelligence and interaction technologies using computer vision, focusing on public spaces such as retail and museums.