Recovering 3D Human Body Configurations Using Shape Contexts

Greg Mori and Jitendra Malik, *Senior Member*, *IEEE*

Abstract—The problem we consider in this paper is to take a single two-dimensional image containing a human figure, locate the joint positions, and use these to estimate the body configuration and pose in three-dimensional space. The basic approach is to store a number of exemplar 2D views of the human body in a variety of different configurations and viewpoints with respect to the camera. On each of these stored views, the locations of the body joints (left elbow, right knee, etc.) are manually marked and labeled for future use. The input image is then matched to each stored view, using the technique of shape context matching in conjunction with a kinematic chain-based deformation model. Assuming that there is a stored view sufficiently similar in configuration and pose, the correspondence process will succeed. The locations of the body joints are then transferred from the exemplar view to the test shape. Given the 2D joint locations, the 3D body configuration and pose are then estimated using an existing algorithm. We can apply this technique to video by treating each frame independently—tracking just becomes repeated recognition. We present results on a variety of data sets.

Index Terms—Shape, object recognition, tracking, human body pose estimation.

1 Introduction

As indicated in Fig. 1, the problem we consider in this paper is to take a single two-dimensional image containing a human figure, locate the joint positions, and use these to estimate the body configuration and pose in three-dimensional space. Variants include the case of multiple cameras viewing the same human, tracking the body configuration and pose over time from video input, or analogous problems for other articulated objects such as hands, animals, or robots. A robust, accurate solution would facilitate many different practical applications—e.g., see Table 1 in Gavrila's survey paper [1]. From the perspective of computer vision theory, this problem offers an opportunity to explore a number of different trade-offs—the role of low-level versus high level cues, static versus dynamic information, 2D versus 3D analysis, etc., in a concrete setting where it is relatively easy to quantify success or failure.

In this paper, we consider the most basic version of the problem—estimating the 3D body configuration based on a single uncalibrated 2D image. The approach we use is to store a number of exemplar 2D views of the human body in a variety of different configurations and viewpoints with respect to the camera. On each of these stored views, the locations of the body joints (left elbow, right knee, etc.) are manually marked and labeled for future use. The test image is then matched to each stored view, using the shape context matching technique of Belongie et al. [2]. This technique is based on representing a shape by a set of sample points from the external and internal contours of an object, found using an edge detector. Assuming that there is a stored view sufficiently similar in configuration and pose, the correspondence process will succeed. The locations of the body joints are then transferred from the exemplar view to the test shape. Given the 2D joint locations, the 3D body configuration and pose are estimated using the algorithm of Taylor [3].

The main contribution of this work is demonstrating the use of deformable template matching to exemplars as a means to localize human body joint positions. Having the context of the whole body, from exemplar templates, provides a wealth of information for matching. The major issue that must be addressed with this approach is dealing with the large number of exemplars needed to match people in a wide range of poses, viewed from a variety of camera positions, and wearing different clothing. In our work, we represent exemplars as a collection of edges extracted using an edge detector and match based on shape in order to reduce the effects of variation in appearance due to clothing. Pose variation presents an immense challenge. In this work, we do not attempt to estimate joint locations for people in arbitrary poses, instead restricting ourselves to settings in which the set of poses is limited (e.g., walking people or speed skaters). Even in such settings, the number of exemplars needed can be very large. In this work, we also provide a method for efficiently retrieving from a large set of exemplars those which are most similar to a query image, in order to reduce the computational expense of matching.

The structure of this paper is as follows: We review previous work in Section 2. In Section 3, we describe the correspondence process mentioned above. We give an efficient method for scaling to large sets of exemplars in Section 4. Section 5 provides details on a parts-based extension to our keypoint estimation method. We describe the 3D estimation algorithm in Section 6. We show experimental results in Section 7. Finally, we conclude in Section 8.

2 Previous Work

There has been considerable previous work on this problem [1]. Broadly speaking, it can be categorized into two major classes. The first set of approaches use a 3D model for

- *G. Mori is with the School of Computing Science, Simon Fraser University, Burnaby, BC V5A 1S6 Canada. E-mail: mori@cs.sfu.ca.*
- *J. Malik is with the Computer Science Division, Department of Electrical Engineering and Computer Science, University of California at Berkeley, Berkeley, CA 94720-1776. E-mail: malik@cs.berkeley.edu.*

Manuscript received 18 Oct. 2004; revised 3 Oct. 2005; accepted 11 Oct. 2005; published online 11 May 2006.
Recommended for acceptance by P. Fua.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMI-0564-1004.

Published by the IEEE Computer Society

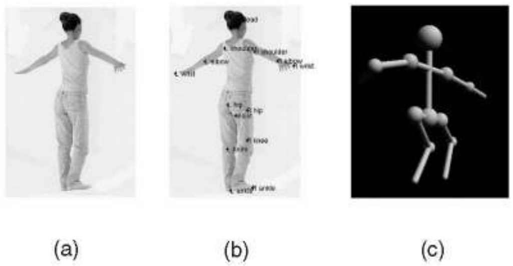

Fig. 1. The goal of this work. (a) Input image. (b) Automatically extracted keypoints. (c) Three-dimensional rendering of estimated body configuration. In this paper, we present a method to go from (a) to (b) to (c).

estimating the positions of articulated objects. Pioneering work was done by O'Rourke and Badler [4], Hogg [5], and Yamamoto and Koshikawa [6]. Rehg and Kanade [7] track very high DOF articulated objects such as hands. Bregler and Malik [8] use optical flow measurements from a video sequence to track joint angles of a 3D model of a human, using the product of exponentials representation for the kinematic chain. Kakadiaris and Metaxas [9] use multiple cameras and match occluding contours with projections from a deformable 3D model. Gavrila and Davis [10] is another 3D model-based tracking approach, as is the work of Rohr [11] for tracking walking pedestrians. Sidenbladh and Black [12] presented a learning approach for developing the edge cues typically used when matching the 3D models projected into the image plane. The method first learns the appearance of edge cues on human figures from a collection of training images and then uses these learned statistics to track people in video sequences. Attempts have also been made at addressing the high-dimensional, multimodal nature of the search space for a 3D human body model. Deutscher et al. [13] have tracked people performing varied and atypical actions using improvements on a particle filter. Choo and Fleet [14] use a Hybrid Monte Carlo (HMC) filter, which at each time step runs a collection of Markov Chain Monte Carlo (MCMC) simulations initialized using a particle filtering approach. Sminchisescu and Triggs [15] use a modified MCMC algorithm to explore the multiple local minima inherent in fitting a 3D model to given 2D image positions of joints. Lee and Cohen [16] presented impressive results on automatic pose estimation from a single image. Their method used *proposal maps*, based on face and skin detection, to guide a MCMC sampler to promising regions of the image when fitting a 3D body model.

The second broad class of approaches does not explicitly work with a 3D model, rather 2D models trained directly from example images are used. There are several variations on this theme. Baumberg and Hogg [17] use active shape models to track pedestrians. Wren et al. [18] track people as a set of colored blobs. Morris and Rehg [19] describe a 2D scaled prismatic model for human body registration. Ioffe and Forsyth [20] perform low-level processing to obtain candidate body parts and then use a mixture of trees to infer likely configurations. Ramanan and Forsyth [21] use similar low-level processing, but add a constraint of temporal appearance consistency to track people and animals in video sequences. Song et al. [22] also perform inference on a tree model, using extracted point features along with motion information. Brand [23] learns a probability distribution over pose and velocity configurations of the moving body and uses it to infer paths in this space. Toyama and Blake [24] use 2D exemplars, scored by comparing edges with Chamfer matching, to track people in video sequences. Most related to our method is the work of Sullivan and Carlsson [25], who use *order structure* to compare exemplar shapes with test images. This approach was developed at the same time as our initial work using exemplars [26].

Other approaches rely on background subtraction to extract a silhouette of the human figure. A mapping from silhouettes to 3D body poses is learned from training images, and applied to the extracted silhouettes to recover pose. Rosales and Sclaroff [27] describe the Specialized Mappings Architecture (SMA), which incorporates the inverse 3D pose to silhouette mapping for performing inference. Grauman et al. [28] learn silhouette contour models from multiple cameras using a large training set obtained by rendering synthetic human models in a variety of poses. Haritaoglu et al. [29] first estimate approximate posture of the human figure by matching to a set of prototypes. Joint positions are then localized by finding extrema and curvature maxima on the silhouette boundary.

Our method first localizes joint positions in 2D and then lifts them to 3D using the geometric method of Taylor [3]. There are a variety of alternative approaches to this lifting problem. Lee and Chen [30], [31] preserve the ambiguity regarding foreshortening (closer endpoint of each link) in an interpretation tree and use various constraints to prune impossible configurations. Attwood et al. [32] use a similar formulation and evaluate the likelihood of interpretations based on joint angle probabilities for known posture types. Ambrósio et al. [33] describe a photogrammetric approach that enforces temporal smoothness to resolve the ambiguity due to foreshortening. Barrón and Kakadiaris [34] simultaneously estimate 3D pose and anthropometry (body parameters) from 2D joint positions in a constrained optimization method.

3 Estimation Method

In this section, we provide the details of the configuration estimation method proposed above. We first obtain a set of boundary sample points from the image. Next, we estimate the 2D image positions of 14 *keypoints* (wrists, elbows, shoulders, hips, knees, ankles, head, and waist) on the image by deformable matching to a set of stored exemplars that have hand-labeled keypoint locations. These estimated keypoints can then be used to construct an estimate of the 3D body configuration in the test image.

3.1 Deformable Matching Using Shape Contexts

Given an exemplar (with labeled keypoints) and a test image, we cast the problem of keypoint estimation in the test image as one of deformable matching. We attempt to deform the exemplar (along with its keypoints) into the shape of the test image. Along with the deformation, we compute a matching score to measure similarity between the deformed exemplar and the test image.

In our approach, a shape is represented by a discrete set of n points $\mathcal{P} = \{p_1, \dots, p_n\}$, $p_i \in \mathbb{R}^2$ sampled from the internal and external contours on the shape.

We first perform edge detection on the image, using the boundary detector of Martin et al. [35], to obtain a set of edge pixels on the contours of the body. We then sample some number of points (300-1,000 in our experiments) from these

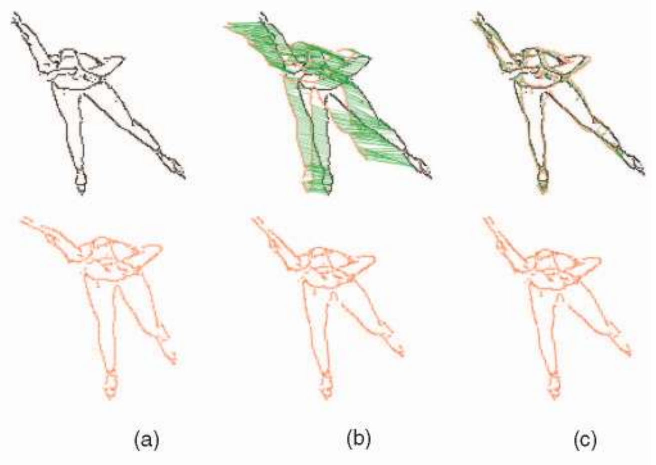

Fig. 2. Iterations of deformable matching. (a) Shows sample points from the two figures to be matched. The bottom figure (exemplar) in (a) is deformed into the shape of the top figure (test image). (b) and (c) Show successive iterations of deformable matching. The top row shows the correspondences obtained through the shape context matching. The bottom row shows the deformed exemplar figure at each step. In particular, the right arm and left leg of the exemplar are deformed into alignment with the test image.

edge pixels to use as the sample points for the body. Note that this process will give us not only external, but also internal contours of the body shape. The internal contours are essential for estimating configurations of self-occluding bodies.

The deformable matching process consists of three steps. Given sample points on the exemplar and test image:

1. Obtain correspondences between exemplar and test image sample points.
2. Estimate deformation of exemplar.
3. Apply deformation to exemplar sample points.

We perform a small number (maximum of four in experiments) of iterations of this process to match an exemplar to a test image. Fig. 2 illustrates this process.

3.1.1 Sample Point Correspondences

In the correspondence phase, for each point p_i on a given shape, we want to find the "best" matching point q_j on another shape. This is a correspondence problem similar to that in stereopsis. Experience there suggests that matching is easier if one uses a rich local descriptor. Rich descriptors reduce the ambiguity in matching.

The *shape context* was introduced by Belongie et al. [2] to play such a role in shape matching. In later work [36], we extended the shape context descriptor by encoding more descriptive information than point counts in the histogram bins. To each edge point q_j, we attach a unit length tangent vector t_j that is the direction of the edge at q_j. In each bin, we sum the tangent vectors for all points falling in the bin. The descriptor for a point p_i is the histogram $\hat{h}_i$:

$$\hat{h}_i^k = \sum_{q_j \in Q} t_j, \text{where } Q = \{q_j \neq p_i, (q_j - p_i) \in \text{bin}(k)\}. \quad (1)$$

Each histogram bin $\hat{h}_i^k$ now holds a single vector in the direction of the dominant orientation of edges falling in the spatial area $\text{bin}(k)$. When comparing the descriptors for two points, we convert this d-bin histogram to a $2D\text{-dimensional}$ vector $\hat{v}_i$, normalize these vectors, and compare them using the L^2 norm.

$$\hat{v}_i = \langle \hat{h}_i^{1,x}, \hat{h}_i^{1,y}, \hat{h}_i^{2,x}, \hat{h}_i^{2,y}, \ldots, \hat{h}_i^{d,x}, \hat{h}_i^{d,y} \rangle, \quad (2)$$

where $\hat{h}_i^{j,x}$ and $\hat{h}_i^{j,y}$ are the x and y components of $\hat{h}_i^j$, respectively.

We call these extended descriptors *generalized shape contexts*. Examples of these generalized shape contexts are shown in Fig. 3. Note that generalized shape contexts reduce to the original shape contexts if all tangent angles are clamped to zero. As in the original shape contexts, these descriptors are not scale invariant. In the absence of substantial background clutter, scale invariance can be achieved by setting the bin radii as a function of average interpoint distances. Some amount of rotational invariance is obtained via the binning structure, as after a small rotation sample points will still fall in the same bins. Full rotational invariance can be obtained by fixing the orientation of the histograms with respect to a local edge tangent estimate. In this work, we do not use these strategies for full scale and rotational invariance. This has the drawback of possibly requiring more exemplars. However, there are definite advantages. For example, people tend to appear in upright poses. By not having a descriptor with full rotational invariance, we are very unlikely to confuse sample points on the feet with those on the head.

We desire a correspondence between sample points on the two shapes that enforces the uniqueness of matches. This leads us to formulate our matching of a test image to an exemplar human figure as an assignment problem (also known as the weighted bipartite matching problem) [37]. We find an optimal assignment between sample points on the test body and those on the exemplar.

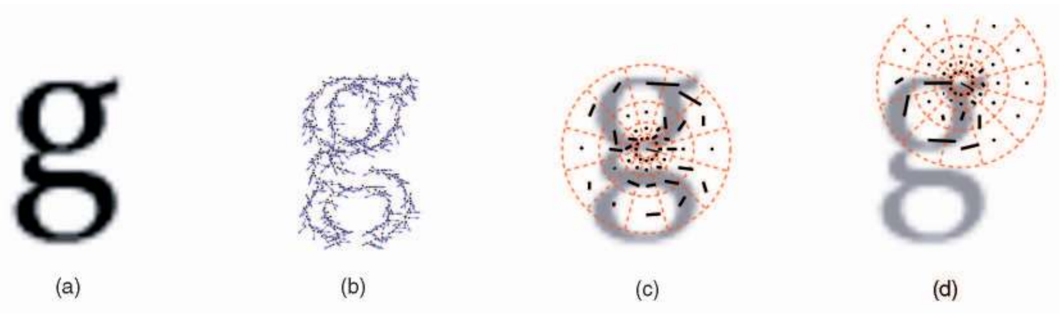

Fig. 3. Examples of generalized shape contexts. (a) Input image. (b) Sampled edge point with tangents. (c) and (d) Generalized shape contexts for different points on the shape.

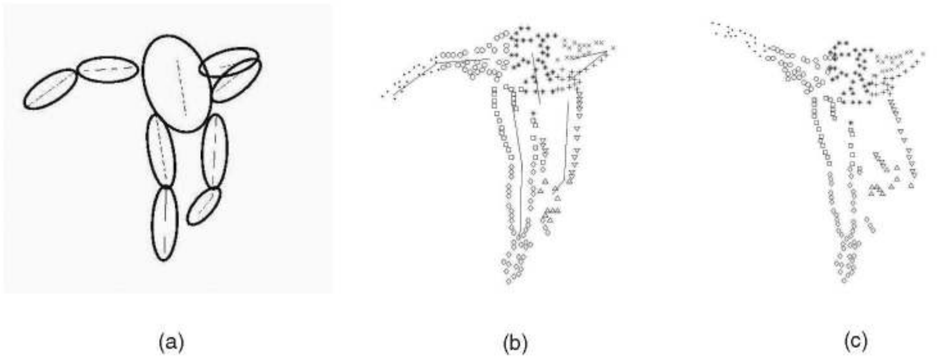

Fig. 4. The deformation model. (a) Underlying kinematic chain. (b) Automatic assignment of sample points to kinematic chain segments on an exemplar. Each different symbol denotes a different chain segment. (c) Sample points deformed using the kinematic chain.

To this end, we construct a bipartite graph. The nodes on one side represent sample points from the test image, on the other side the sample points on the exemplar. Edge weights between nodes in this bipartite graph represent the costs of matching sample points. Similar sample points will have a low matching cost, dissimilar ones will have a high matching cost. ϵ-cost outlier nodes are added to the graph to account for occluded points and noise—sample points missing from a shape can be assigned to be outliers for some small cost. We use an assignment problem solver to find the optimal matching between the sample points of the two bodies.

Note that the output of more specific filters, such as face or hand detectors, could easily be incorporated into this framework. The matching cost between sample points can be measured in many ways.

3.1.2 Deformation Model

Belongie et al. [2] used thin plate splines as a deformation model. However, it is not appropriate here, as human figures deform in a more structured manner. We use a 2D kinematic chain as our deformation model. The 2D kinematic chain has nine segments: a torso (containing head, waist, hips, and shoulders), upper and lower arms (linking elbows to shoulders, and wrists to elbows), and upper and lower legs (linking knees to hips, and ankles to knees). Fig. 4a depicts the kinematic chain deformation model. Our deformation model allows translation of the torso and 2D rotation of the limbs around the shoulders, elbows, hips, and knees. This is a simple representation for deformations of a figure in 2D. It only allows in-plane rotations, ignoring the effects of perspective projection as well as out of plane rotations. However, this deformation model is sufficient to allow for small deformations of an exemplar.

In order to estimate a deformation or deform a body's sample points, we must know to which kinematic chain segment each sample point belongs. On the exemplars, we have hand-labeled keypoints; we use these to automatically assign the hundreds of sample points to segments. Sample points are assigned to segments by finding minimum distance to bone-line, the line segment connecting the keypoints at the segment ends, for arm and leg segments. For the torso, line segments connecting the shoulders and hips are used. A sample point is assigned to the segment for which this distance is smallest.

Since we know the segment $S(p_i)$ that each exemplar sample point p_i belongs to, given correspondences $\{(p_i, p_i')\}$, we can estimate a deformation D of the points $\{p_i\}$. Our deformation process starts at the torso. We find the least squares best translation for the sample points on the torso.

$$D_t = \hat{T} = \arg\min{}_T \sum_{p_i, S(p_i)=torso} \|T(p_i) - {p_i}'\|^2, \qquad (3)$$

$$\hat{T} = \frac{1}{N} \sum_{p_i : S(p_i)=torso} ({p_i}' - p_i), \text{where } N = \#\{p_i : S(p_i) = torso\}. \qquad (4)$$

Subsequent segments along the kinematic chain have rotational joints. We again obtain the least squares best estimates, this time for the rotations of these joints. Given previous deformation $\hat{D}$ along the chain up to this segment, we estimate D_j as the best rotation around the joint location c_j:

$$P_j = \{p_i : S(p_i) = j\}, \qquad (5)$$

$$D_j = R_{\hat{\theta}, c_j} = \arg\min{}_{R_{\theta}, c_j} \sum_{p_i \in P_j} \|R_{\theta, c_j}(\hat{D} \cdot p_i) - {p_i}'\|^2, \qquad (6)$$

$$\hat{\theta} = \arg\min{}_\theta \sum_{p_i \in P_j} (\hat{D} \cdot p_i - c_j)^T R_\theta^T (c_j - p_i'), \qquad (7)$$

$$\hat{\theta} = \arctan \frac{\sum_i q_{ix} q_{iy}' - \sum_i q_{iy} q_{ix}'}{\sum_i q_{ix} q_{ix}' + \sum_i q_{iy} q_{iy}'}, \qquad (8)$$

$$\text{where } q_i = \hat{D} \cdot p_i - c_j \text{ and } q_i' = p_i' - c_j. \qquad (9)$$

Steps 2 and 3 in our deformable matching framework are performed in this manner. We estimate deformations for each segment of our kinematic chain model and apply them to the sample points belonging to each segment.

We have now provided a method for estimating a set of keypoints using a single exemplar, along with an associated score (the sum of shape context matching costs for the optimal assignment). The simplest method for choosing the best keypoint configuration in a test image is to find the exemplar with the best score and use the keypoints predicted using its deformation as the estimated configuration. However, with this simple method, there are concerns involving the number of exemplars needed for a general matching framework. In the following sections, we will address this by first describing an efficient method for scaling to large sets of exemplars and then developing a parts-based method for combining matching results from multiple exemplars.

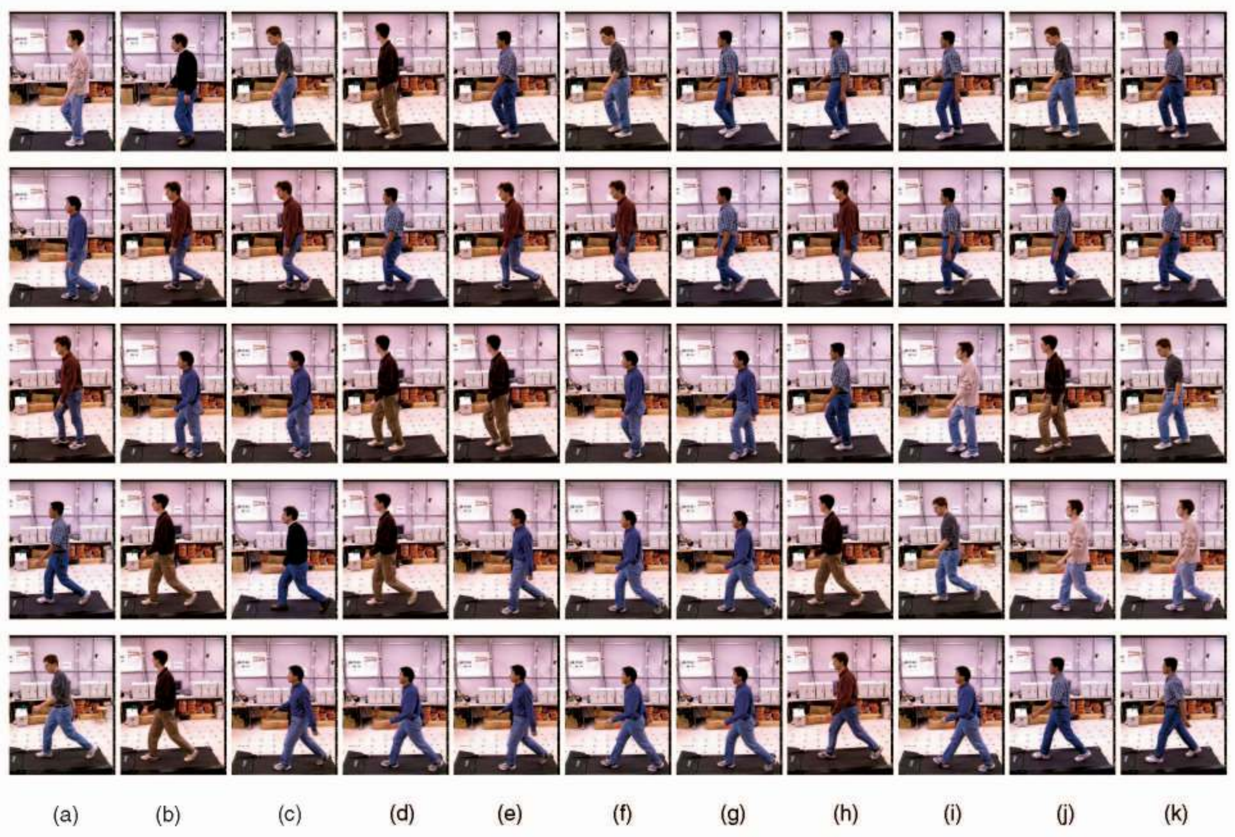

Fig. 5. Example shortlists. (a) Shows query image. (b), (c), (d), (e), (f), (g), (h), (i), (j), and (k) Show shortlist of candidate matches from representative shape context pruning. Exemplars in poses similar to the human figure in the query image are retrieved.

4 Scaling to Large Sets of Exemplars

The deformable matching process described above is computationally expensive. If we have a large set of exemplars, which will be necessary in order to match people of different body shapes in varying poses, performing an exhaustive comparison to every exemplar is not feasible. Instead, we use an efficient pruning algorithm to reduce the full set of exemplars to a shortlist of promising candidates. Only this small set of candidates will be compared to the test image using the expensive deformable matching process.

In particular, we use the *representative shape contexts* pruning algorithm [38] to construct this shortlist of candidate exemplars. This method relies on the descriptive power of just a few shape contexts. Given a pair of images of very different human figures, such as a tall person walking and a short person jogging, none of the shape contexts from the walking person will have good matches on the jogging one—it is immediately obvious that they are different shapes. The representative shape contexts pruning algorithm uses this intuition to efficiently construct a shortlist of candidate matches.

In concrete terms, the pruning process proceeds in the following manner. For each of the exemplar human figure shapes S_i, we precompute a large number s (about 800) of shape contexts $\{SC_i^j : j = 1, 2, \dots, s\}$. But for the query human figure shape S_q, we only compute a small number r ($r \approx 5 - 10$ in experiments) of representative shape contexts (RSCs). To compute these r RSCs, we randomly select r sample points from the shape via a rejection sampling method that spreads the points over the entire shape. We use all the sample points on the shape to fill the histogram bins for the shape contexts corresponding to these r points. To compute the distance between a query shape and an exemplar shape, we find the best matches for each of the r RSCs.

The distance between shapes S_q and S_i is then:

$$d_S(S_q, S_i) = \frac{1}{r}\sum_{u=1}^{r} \frac{d_{GSC}(SC_q^u, SC_i^{m(u)})}{N_u}, \tag{10}$$

$$\text{where } m(u) = \arg\min_j d_{GSC}(SC_q^u, SC_i^j). \tag{11}$$

N_u is a normalizing factor that measures how discriminative the representative shape context SC_q^u is:

$$N_u = \frac{1}{|\mathbb{S}|}\sum_{S_i \in \mathbb{S}} d_{GSC}(SC_q^u, SC_i^{m(u)}), \tag{12}$$

where $\mathbb{S}$ is the set of all shapes. We determine the shortlist by sorting these distances. Fig. 5 shows some example shortlists. Note that this pruning method, as presented, assumes that the human figure is the only object in the query image, as will be the case in our experiments. However, it is possible to run this pruning method in cluttered images [38].

5 Using Part Exemplars

Given a set of exemplars, we can choose to match either entire exemplars or parts, such as limbs, to a test image. The advantage of a parts-based approach that matches limbs is that of compositionality, which saves us from an exponential explosion in the required number of exemplars. Consider the case of a person walking while holding a briefcase in one hand. If we already have exemplars for a walking motion and a single exemplar for holding an object in the hand, we can combine these exemplars to produce correct matching results. However, if we were forced to use entire exemplars, we would require a different "holding object and walking" exemplar for each portion of the walk cycle. Using part exemplars prevents the total number of exemplars from growing to an unwieldy size. As long as we can ensure that the composition of part exemplars yields an anatomically correct configuration, we will benefit from this reduced number of exemplars.

The matching process is identical to that presented in the preceding section. For each exemplar, we deform it to the shape of the test image. However, instead of assigning a total score for an exemplar, we give a separate score for each part on the exemplar. This is done by summing the shape context matching costs for sample points from each part. In our experiments (Fig. 8), we use six "limbs" as our parts: arms (consisting of shoulder, elbow, and wrist keypoints) and legs (hip, knee, and ankle), along with separate head and waist parts.

With N exemplars we have N estimates for the location of each of the six limbs. Each of these N estimates is obtained using the deformable matching process described in the previous section. We will denote by l_i^j the $j\mathrm{th}$ limb obtained by matching to the $i\mathrm{th}$ exemplar, and its shape context matching score (obtained from the deformable matching process) to be L_i^j. We now combine these individual matching results to find the "best" combination of these estimates. It is not sufficient to simply choose each limb independently as the one with the best score. There would be nothing to prevent us from violating underlying anatomical constraints. For example, the left leg could be found hovering across the image disjoint from the rest of the body. We need to enforce the *consistency* of the final configuration.

Consider again the case of using part exemplars to match the figure of a person walking while holding a briefcase. Given a match for the arm grasping the briefcase and matches for the rest of the body, we know that there are constraints on the distance between the shoulder of the grasping arm and the rest of the body. Motivated by this, the measure of consistency we use is the 2D image distance between the bases (shoulder for the arms, hip for the legs) of limbs. We form a tree structure by connecting the arms and the waist to the head and the legs to the waist. For each link in this tree, we compute the N^2 2D image distances between all pairs of bases of limbs obtained by matching with the N different exemplars. We now make use of the fact that each whole exemplar on its own is consistent. Consider a pair of limbs (l_i^u, l_j^v)—limb u from exemplar i and limb v from exemplar j, with (u, v) being a link in the tree, such as left hip—waist. Using the limbs from these two different exemplars together is plausible if the distances between their bases is comparable to that of each of the whole exemplars. We compare the distance d_{ij}^{uv} between the bases b_i^u and b_j^v of these limbs with the two distances obtained when taking limbs u and v to be both from exemplar i or both from exemplar j. We define the consistency cost C_{ij}^{uv} of using this pair of limbs (l_i^u, l_j^v) together in matching a test image to be a function of the average of the two differences, scaled by a parameter σ:

$$d_{ij}^{uv} = \|b_i^u - b_j^v\|, \tag{13}$$

$$C_{ij}^{uv} = 1 - exp\left(-\frac{|d_{ij}^{uv} - d_{ii}^{uv}| + |d_{ij}^{uv} - d_{jj}^{uv}|}{2\sigma}\right). \tag{14}$$

Note that the consistency cost C_{ii}^{uv} for using limbs from the same exemplar across a tree link is zero. As the configuration begins to deviate from the consistent exemplars, C_{ij}^{uv} increases. We define the total cost $S(x)$ of a configuration $x = (x^1, x^2, \dots, x^6) \in \{1, 2, \dots, N\}^6$ as the weighted sum of consistency scores and shape context limb scores $L_{x^j}^j$:

$$S(x) = (1 - w_c)\sum_{j=1}^{6} L_{x^j}^j + w_c \sum_{links:(i,j)} C_{x^i x^j}^{ij}. \tag{15}$$

The relative importance between quality of individual scores and consistency costs is determined by w_c. Both w_c and σ (defined above) were determined manually. Note that, when using part exemplars, shape contexts are still computed using sample points from whole exemplars. In our experiments, we did not find the use of shape context limb scores from whole exemplars to be problematic, possibly due to the coarse binning structure of the shape contexts.

There are N^6 possible combinations of limbs from the N exemplars. However, we can find the optimal configuration in $O(N^2)$ time using a dynamic programming algorithm along the tree structure.

Moreover, an extension to our algorithm can produce the top K matches for a given test image. Preserving the ambiguity in this form, instead of making an instant choice, is particularly advantageous for tracking applications, where temporal consistency can be used as an additional filter.

6 Estimating 3D Configuration

We use Taylor's method [3] to estimate the 3D configuration of a body given the keypoint position estimates. Taylor's method works on a single 2D image, taken with an uncalibrated camera.

It assumes that we know:

1. the image coordinates of keypoints (u, v),
2. the relative lengths l of body segments connecting these keypoints,
3. a labeling of "closer endpoint" for each of these body segments, and
4. that we are using a scaled orthographic projection model for the camera.

In our work, the image coordinates of keypoints are obtained via the deformable matching process. The "closer endpoint" labels are supplied on the exemplars and automatically transferred to an input image after the matching process. The relative lengths of body segments are fixed in advance, but could also be transferred from exemplars.

We use the same 3D kinematic model defined over keypoints as that in Taylor's work.

We can solve for the 3D configuration of the body $\{(X_i, Y_i, Z_i) : i \in keypoints\}$ up to some ambiguity in scale s. The method considers the foreshortening of each body segment to construct the estimate of body configuration.

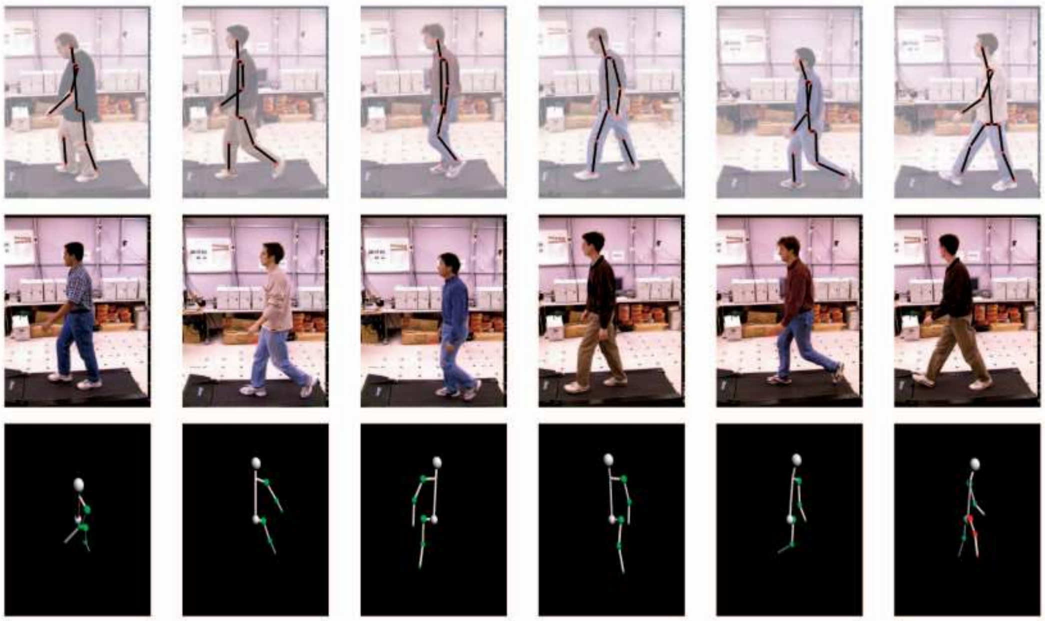

Fig. 6. Results on MoBo data set. The top row shows input image with recovered joint positions. The middle row shows best matching exemplar, from which joint positions were derived. The bottom row shows 3D reconstruction from different viewpoint. Only joint positions marked as unoccluded on the exemplar are transferred to the input image. Joint positions are marked as red dots, black lines connect unoccluded joints adjacent in the body model. Note that background subtraction is performed to remove clutter in this data set.

For each pair of body segment endpoints, we have the following equations:

$$l^2 = (X_1 - X_2)^2 + (Y_1 - Y_2)^2 + (Z_1 - Z_2)^2, \tag{16}$$

$$(u_1 - u_2) = s(X_1 - X_2), \tag{17}$$

$$(v_1 - v_2) = s(Y_1 - Y_2), \tag{18}$$

$$dZ = (Z_1 - Z_2), \tag{19}$$

$$\Longrightarrow dZ = \sqrt{l^2 - ((u_1 - u_2)^2 + (v_1 - v_2)^2)/s^2}. \tag{20}$$

To estimate the configuration of a body, we first fix one keypoint as the reference point and then compute the positions of the others with respect to the reference point. Since we are using a scaled orthographic projection model, the X and Y coordinates are known up to the scale s. All that remains is to compute relative depths of endpoints dZ. We compute the amount of foreshortening and use the user-supplied "closer endpoint" labels from the closest matching exemplar to solve for the relative depths.

Moreover, Taylor notes that the minimum scale s_{min} can be estimated from the fact that dZ cannot be complex.

$$s \geq \frac{\sqrt{(u_1 - u_2)^2 + (v_1 - v_2)^2}}{l}. \tag{21}$$

This minimum value is a good estimate for the scale since one of the body segments is often perpendicular to the viewing direction.

7 Experiments

We demonstrate results of our method applied to three domains—video sequences of walking people from the CMU MoBo Database, a speed skater, and a running cockroach. In all of these video sequences, each frame is processed independently—no dynamics are used and no temporal consistency is enforced.

Each of these experiments presents a challenge in terms of variation in pose within a restricted domain. In the case of the MoBo Database, substantial variation in clothing and body shape are also present. We do not address the problem of background clutter. In each of the data sets, either a simple background exists or background subtraction is used, so that the majority of extracted edges belong to the human figure in the image.

7.1 CMU MoBo Database

The first set of experiments we performed used images from the CMU MoBo Database [39]. This database consists of video sequences of number of subjects, performing different types of walking motions on a treadmill, viewed from a set of stationary cameras. We selected the first 10 subjects (numbers 04002-04071), 30 frames (frames numbered 101-130) from the "fastwalk" sequence for each subject, and a camera view perpendicular to the direction of the subject's walk (vr03_7). Marking of exemplar joint locations, in addition to "closer endpoint" labels, was performed manually on this collection of 300 frames. Background subtraction was used to remove most of the clutter edges found by the edge detector.

We used this data set to study the ability of our method to handle variations in body shape and clothing. A set of 10 experiments was conducted in which each subject was used once as the query against a set of exemplars consisting of the images of the remaining nine subjects. For each query image, this set of 270 exemplars was pruned to a shortlist of length 10 using representative shape contexts. Deformable matching to localize body joints is only performed using this shortlist. In our unoptimized MATLAB implementation, deformable matching between a query and an exemplar takes 20-30 seconds on a 2 GHz AMD Opteron processor. The

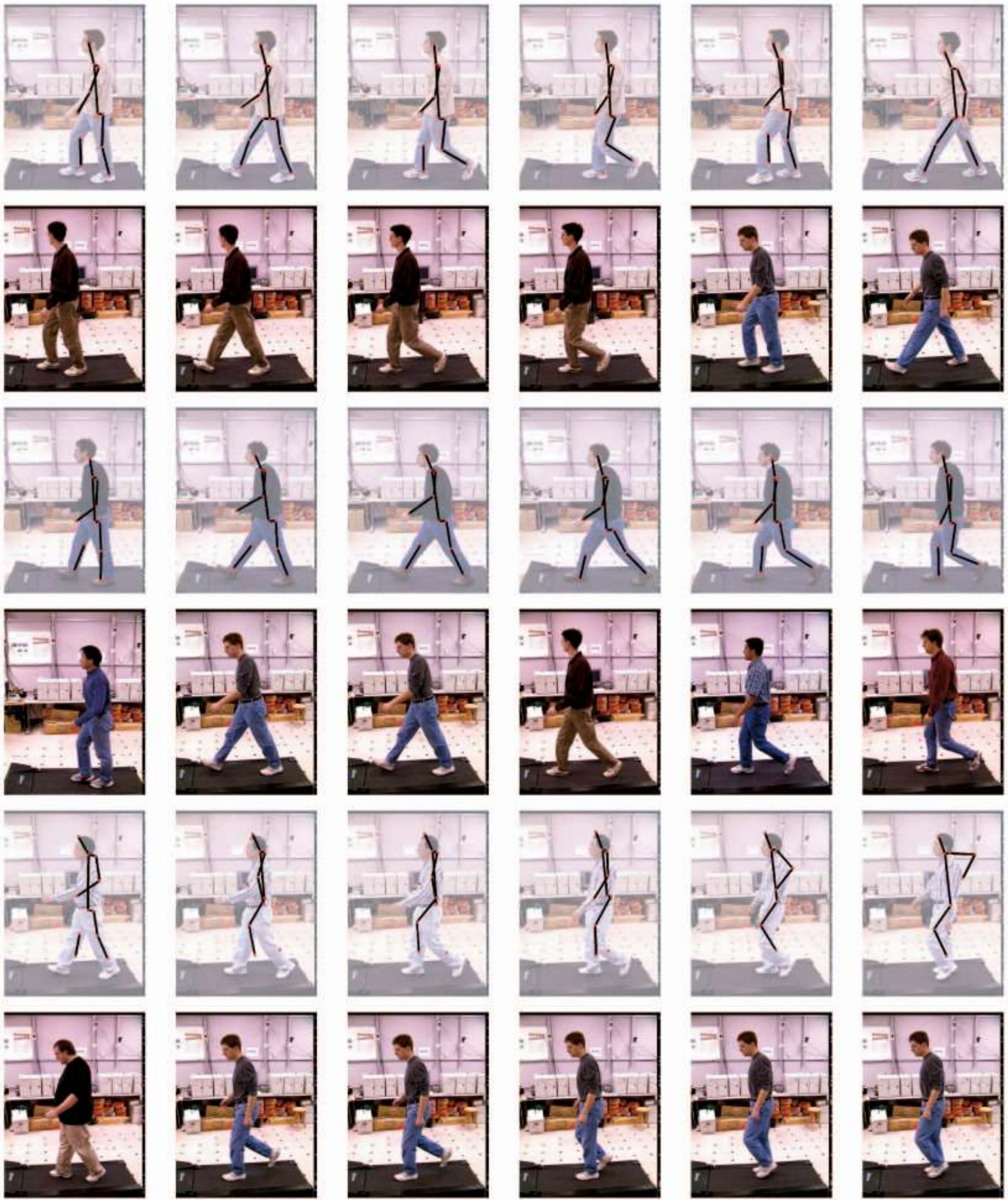

Fig. 7. Results on MoBo data set. Each pair of rows shows input images with recovered joint positions above best matching exemplars. Only joint positions marked as unoccluded on the exemplar are transferred to the input image. Note that background subtraction is performed to remove clutter in this data set.

representative shape contexts pruning takes a fraction of a second and reduces overall computation time substantially.

Note that on this data set keypoints on the subject's right arm and leg are often occluded and are labeled as such. Limbs with occluded joints are not assigned edge points in the deformable matching and, instead, inherit the deformation of limbs further up the kinematic chain. Occluded joints from an exemplar are not transferred onto a query image and are omitted from the 3D reconstruction process.

Fig. 6 shows sample results of 2D body joint localization and 3D reconstruction on the CMU MoBo data set. The same body parameters (lengths of body segments) are used in all 3D reconstructions. With additional manual labeling, these body parameters could be supplied for each exemplar and transferred onto the query image to obtain more accurate reconstructions.

More results of 2D joint localization are shown in Fig. 7. Given good edges, particularly on the subject's arms, the deformable matching process performs well. However, in cases such as the third subject in Fig. 7, the edge detector has difficulty due to clothing. Since the resulting edges are substantially different from those of other subjects, the joint localization process fails.

Fig. 8 shows a comparison between the parts-based dynamic programming approach and single exemplar matching. The parts-based approach is able to improve the localization of joints by combining limbs from different exemplars. The main difficulty encountered with this method is in the reuse of edge pixels. A major source of error is matching the left and right legs of two exemplars to the same edge pixels in the query image. This reuse is a fundamental problem with tree models.

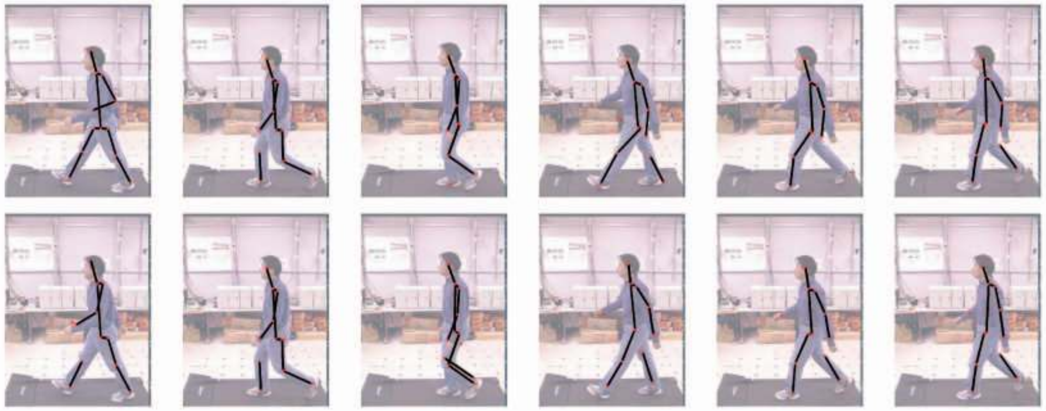

Fig. 8. Comparison between single exemplar and dynamic programming. The top row shows results obtained matching to a single exemplar and the bottom row uses dynamic programming to combine limbs from multiple exemplars. The third column shows an example of reuse of edge pixels to match left and right legs at same location.

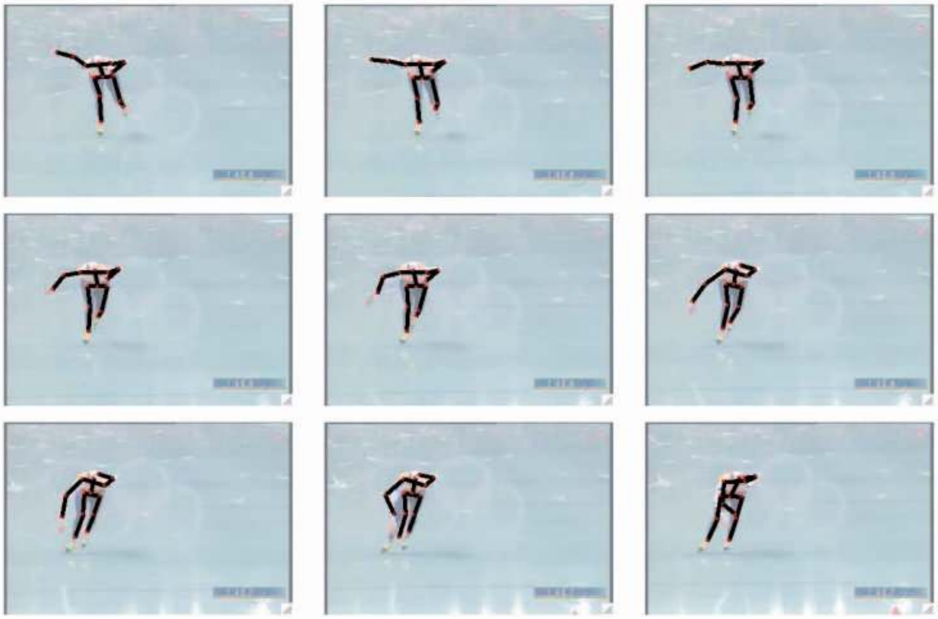

Fig. 9. Results on speed skater sequence. Frames 6-8, 10-12, and 14-16 are shown. Exemplars for the sequence are frames 5, 9, 13, and 17.

7.2 Speed Skating

We also applied our method to a sequence of video frames of a speed skater. We chose five frames for use as exemplars, upon which we hand-labeled keypoint locations. We then applied our method for configuration estimation to a sequence of 20 frames. Results are shown in Fig. 9.

Difficulties are encountered as the skater's arm crosses in front of her body. More exemplars would likely be necessary at these points in the sequence where the relative ordering of edges changes (i.e., furthest left edge is now the edge of thigh instead of the edge of the arm).

7.3 Cockroach Video Sequence

The final data set consisted of 300 frames from a video of a cockroach running on a transparent treadmill apparatus, viewed from below. These data were collected by biologists at UC Berkeley who are studying their movements. The research that they are conducting requires the extraction of 3D joint angle tracks for many hours of footage. The current solution to this tracking problem is manual labor. In each frame of each sequence, a person manually marks the 2D locations of each of the cockroach's joints. 3D locations are typically obtained using stereo from a second, calibrated camera.

Such a setting is ideal for an exemplar-based approach. Even if every 10th frame from a sequence needs to be manually marked and used as an exemplar, a huge gain in efficiency could be made.

As a preliminary attempt at tackling this problem, we applied the same techniques that we developed for detecting human figures to this problem of detecting cockroaches. The method and parameters used were identical, aside from addition of two extra limbs to our model.

We chose 41 frames from the middle 200 frames (every fifth frame) as exemplars to track the remainder of the sequence. Again, each frame was processed independently to show the efficacy of our exemplar-based method. Of course, temporal consistency should be incorporated in developing a final system for tracking.

Fig. 10 shows some results for tracking using the parts-based method. Results are shown for the first 24 frames, outside of the range of the exemplars, which were selected from frames 50 through 250.

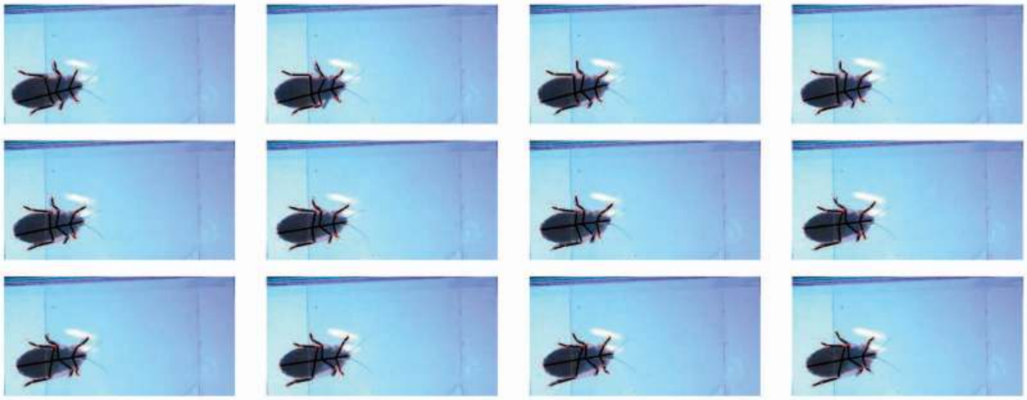

Fig. 10. Results on cockroach sequence. Every second frame of the first 24 frames of the video sequence is shown. The parts-based method was used, with 41 exemplars, every fifth frame starting at frame 50.

8 Conclusion

The problem of recovering human body configurations in a general setting is arguably the most difficult recognition problem in computer vision. By no means do we claim to have solved it here; much work still remains to be done. In this paper, we have presented a simple, yet apparently effective, approach to estimating human body configurations in 3D. Our method matches using 2D exemplars, estimates keypoint locations, and then uses these keypoints in a model-based algorithm for determining the 3D body configuration.

We have shown that using full-body exemplars provides useful context for the task of localizing joint positions. Detecting hands, elbows, or feet in isolation is a difficult problem. A hand is not a hand unless it is connected to an elbow which is connected to a shoulder. Using exemplars captures this type of long-range contextual information. Future work could incorporate additional attributes such as locations of labeled features such as faces or hands in the same framework.

However, there is definitely a price to be paid for using exemplars in this fashion. The number of exemplars needed to match people in a wide range of poses, viewed from a variety of camera positions, is likely to be unwieldy. Recent work by Shakhnarovich et al. [40] has attempted to address this problem of scaling to a large set of exemplars by using locality sensitive hashing to quickly retrieve matching exemplars.

The opposite approach to exemplars of assembling human figures from a collection of low-level parts (e.g., [20], [21], [22], [41]) holds promise in terms of scalability, but, as noted above, lacks the context needed to reliably detect these low-level parts. We believe that combining these two approaches in a sensible manner is an important topic for future work.

Acknowledgments

Portions of this paper appeared in a previous publication [26]. This work was supported by grants from NSERC (RGPIN-312230), the SFU President's Research Fund, and ONR grant N00014-01-1-0890 (MURI). The authors thank the anonymous reviewers for their useful suggestions.

References

[1] D.M. Gavrila, "The Visual Analysis of Human Movement: A Survey," *Computer Vision and Image Understanding: CVIU,* vol. 73, no. 1, pp. 82-98, 1999.

[2] S. Belongie, J. Malik, and J. Puzicha, "Shape Matching and Object Recognition Using Shape Contexts," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 4, pp. 509-522, Apr. 2002.

[3] C.J. Taylor, "Reconstruction of Articulated Objects from Point Correspondences in a Single Uncalibrated Image," *Computer Vision and Image Understanding,* vol. 80, pp. 349-363, 2000.

[4] J. O'Rourke and N. Badler, "Model-Based Image Analysis of Human Motion Using Constraint Propagation," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 2, no. 6, pp. 522-536, 1980.

[5] D. Hogg, "Model-Based Vision: A Program to See a Walking Person," *Image and Vision Computing,* vol. 1, no. 1, pp. 5-20, 1983.

[6] M. Yamamoto and K. Koshikawa, "Human Motion Analysis Based on a Robot Arm Model," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* pp. 664-665, 1991.

[7] J.M. Rehg and T. Kanade, "Visual Tracking of High DOF Articulated Structures: An application to Human Hand Tracking," *Lecture Notes in Computer Science,* vol. 800, pp. 35-46, 1994.

[8] C. Bregler and J. Malik, "Tracking People with Twists and Exponential Maps," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* pp. 8-15, 1998.

[9] I. Kakadiaris and D. Metaxas, "Model-Based Estimation of 3D Human Motion," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 12, pp. 1453-1459, Dec. 2000.

[10] D. Gavrila and L. Davis, "3D Model-Based Tracking of Humans in Action: A MultiView Approach," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* pp. 73-80, 1996.

[11] K. Rohr, "Incremental Recognition of Pedestrians from Image Sequences," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* pp. 8-13, 1993.

[12] H. Sidenbladh and M.J. Black, "Learning the Statistics of People Learning the Statistics of People in Images and Video," *Int'l J. Computer Vision,* vol. 54, nos. 1-3, pp. 183-209, 2003.

[13] J. Deutscher, A.J. Davison, and I.D. Reid, "Automatic Partitioning of High Dimensional Search Spaces Associated with Articulated Body Motion Capture," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 669-676, Dec. 2001.

[14] K. Choo and D.J. Fleet, "People Tracking Using Hybrid Monte Carlo Filtering," *Proc. Eighth Int'l Conf. Computer Vision,* vol. 2, pp. 321-328, 2001.

[15] C. Sminchisescu and B. Triggs, "Hyperdynamic Importance Sampling," *Proc. European Conf. Computer Vision,* vol. 1, pp. 769-783, 2002.

[16] M.W. Lee and I. Cohen, "Proposal Maps Driven MCMC for Estimating Human Body Pose in Static Images," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 334-341, 2004.

[17] A. Baumberg and D. Hogg, "Learning Flexible Models from Image Sequences," *Lecture Notes in Computer Science,* vol. 800, pp. 299-308, 1994.

[18] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-Time Tracking of the Human Body," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 19, no. 7, pp. 780-785, July 1997.

[19] D. Morris and J. Rehg, "Singularity Analysis for Articulated Object Tracking," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* pp. 289-296, 1998.

[20] S. Ioffe and D. Forsyth, "Human Tracking with Mixtures of Trees," *Proc. Eighth Int'l Conf. Computer Vision,* vol. 1, pp. 690-695, 2001.

[21] D. Ramanan and D.A. Forsyth, "Using Temporal Coherence to Build Models of Animals," *Proc. Ninth Int'l Conf. Computer Vision,* vol. 1, pp. 338-345, 2003.

[22] Y. Song, L. Goncalves, and P. Perona, "Unsupervised Learning of Human Motion," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 25, no. 7, pp. 814-827, July 2003.

[23] M. Brand, "Shadow Puppetry," *Proc. Seventh Int'l Conf. Computer Vision,* vol. 2, pp. 1237-1244, 1999.

[24] K. Toyama and A. Blake, "Probabilistic Exemplar-Based Tracking in a Metric Space," *Proc. Eighth Int'l Conf. Computer Vision,* vol. 2, pp. 50-57, 2001.

[25] J. Sullivan and S. Carlsson, "Recognizing and Tracking Human Action," *Proc. European Conf. Computer Vision,* vol. 1, pp. 629-644, 2002.

[26] G. Mori and J. Malik, "Estimating Human Body Configurations Using Shape Context Matching," *Proc. European Conf. Computer Vision,* vol. 3, pp. 666-680, 2002.

[27] R. Rosales and S. Sclaroff, "Learning Body Pose via Specialized Maps," *Neural Information Processing Systems NIPS-14,* 2002.

[28] K. Grauman, G. Shakhnarovich, and T. Darrell, "Inferring 3D Structure with a Statistical Image-Based Shape Model," *Proc. Ninth Int'l Conf. Computer Vision,* 2003.

[29] I. Haritaoglu, D. Harwood, and L.S. Davis, "Ghost: A Human Body Part Labeling System Using Silhouettes," *Proc. Int'l Conf. Pattern Recognition,* 1998.

[30] H.J. Lee and Z. Chen, "Determination of 3D Human Body Posture from a Single View," *Proc. Computer Vision, Graphics, Image Processing,* vol. 30, pp. 148-168, 1985.

[31] Z. Chen and H.J. Lee, "Knowledge-Guided Visual Perception of 3-D Human Gait from a Single Image Sequence," *IEEE Trans. Systems, Man, and Cybernetics,* vol. 22, no. 2, pp. 336-342, 1992.

[32] C.I. Attwood, G.D. Sullivan, and K.D. Baker, "Model-Based Recognition of Human Posture Using Single Synthetic Images," *Proc. Fifth Alvey Vision Conf.,* 1989.

[33] J. Ambrósio, J. Abrantes, and G. Lopes, "Spatial Reconstruction of Human Motion by Means of a Single Camera and a Biomechanical Model," *Human Movement Science,* vol. 20, pp. 829-851, 2001.

[34] C. Barrón and I.A. Kakadiaris, "Estimating Anthropometry and Pose from a Single Uncalibrated Image," *Proc. Computer Vision and Image Understanding (Computer Vision and Image Understanding),* vol. 81, pp. 269-284, 2001.

[35] D. Martin, C. Fowlkes, and J. Malik, "Learning to Find Brightness and Texture Boundaries in Natural Images," *Neural Information Processing Systems,* 2002.

[36] G. Mori and J. Malik, "Recognizing Objects in Adversarial Clutter: Breaking a Visual Captcha," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* vol. 1, pp. 134-141, 2003.

[37] T. Cormen, C. Leiserson, and R. Rivest, *Introduction to Algorithms.* The MIT Press, 1990.

[38] G. Mori, S. Belongie, and J. Malik, "Efficient Shape Matching Using Shape Contexts," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 27, no. 11, pp. 1832-1837, Nov. 2005.

[39] R. Gross and J. Shi, "The CMU Motion of Body (MoBo) Database," Technical Report CMU-RI-TR-01-18, Robotics Inst., Carnegie Mellon Univ., 2001.

[40] G. Shakhnarovich, P. Viola, and T. Darrell, "Fast Pose Estimation with Parameter Sensitive Hashing," *Proc. Ninth Int'l Conf. Computer Vision,* vol. 2, pp. 750-757, 2003.

[41] G. Mori, X. Ren, A. Efros, and J. Malik, "Recovering Human Body Configurations: Combining Segmentation and Recognition," *Proc. IEEE CS Conf. Computer Vision and Pattern Recognition,* vol. 2, pp. 326-333, 2004.

Greg Mori received the Hon BSc degree with high distinction in computer science and mathematics from the University of Toronto in 1999, and the PhD degree in computer science from the University of California at Berkeley in 2004. At Berkeley, he was supported in part by a UC Berkeley Regents' Fellowship. He is currently an assistant professor in the School of Computing Science at Simon Fraser University. His research interests are in computer vision, and include human body pose estimation, activity recognition, shape matching, and object recognition.

Jitendra Malik received the BTech degree in electrical engineering from the Indian Institute of Technology, Kanpur, in 1980 and the PhD degree in computer science from Stanford University in 1985. In January 1986, he joined the University of California at Berkeley, where he is currently the Arthur J. Chick Professor and chair of the Department of Electrical Engineering and Computer Science. He is also on the faculty of the Cognitive Science and Vision Science groups. His research interests are in computer vision and computational modeling of human vision. His work spans a range of topics in vision including image segmentation and grouping, texture, stereopsis, object recognition, image-based modeling and rendering, content-based image querying, and intelligent vehicle highway systems. He has authored or coauthored more than 100 research papers on these topics. He received the gold medal for the best graduating student in electrical engineering from IIT Kanpur in 1980, a Presidential Young Investigator Award in 1989, and the Rosenbaum fellowship for the Computer Vision Programme at the Newton Institute of Mathematical Sciences, University of Cambridge in 1993. He received the Diane S. McEntyre Award for Excellence in Teaching from the Computer Science Division, University of California at Berkeley, in 2000. He was awarded a Miller Research Professorship in 2001. He serves on the editorial board of the *International Journal of Computer Vision*. He is a senior member of the IEEE.