Human Pose Estimation using Motion Exemplars

Alireza Fathi and Greg Mori
School of Computing Science
Simon Fraser University, Burnaby, BC, V5A 1S6 Canada
{alirezaf,mori}@cs.sfu.ca

Abstract

We present a motion exemplar approach for finding body configuration in monocular videos. A motion correlation technique is employed to measure the motion similarity at various space-time locations between the input video and stored video templates. These observations are used to predict the conditional state distributions of exemplars and joint positions. Exemplar sequence selection and joint position estimation are then solved with approximate inference using Gibbs sampling and gradient ascent. The presented approach is able to find joint positions accurately for people with textured clothing. Results are presented on a dataset containing slow, fast and incline walk videos of various people from different view angles. The results demonstrate an overall improvement compared to previous methods.

1. Introduction

In this paper we explore the problem of estimating the pose of a human figure from monocular image sequences. Many practical applications would be enabled by a solution to this problem, including human-computer interaction, gait analysis, and video motion capture. As such it has received a large amount of attention from the computer vision community.

We develop a novel motion-exemplar approach for automatically detecting and tracking human figures in this paper. In our approach we assume we are given a set of exemplar image sequences upon which we have labeled positions of body joints. Given an input image sequence, we infer the pose of the human figure by first finding a sequence of exemplars which match the input sequence, and then estimating body joint positions using these exemplars. Both of these are accomplished by comparing motion estimates for the input sequence against those in the exemplar sequences. Figure 1 shows an overview of our approach.

At the core of most previous approaches to this problem lies a matching of either silhouette (e.g. [13, 17, 1] or edge (e.g. [12, 21, 20, 8, 14]) features for human pose estimation. Compared to these features, the use of motion estimates as a cue has significant advantages.

Approaches which use 2d silhouettes are unable to observe human body limbs when they are in front of the body. In many common poses, the projection of the human figure to the image plane will lead to highly ambiguous 2d silhouette data. Given these ambiguous data as input, pose estimation and tracking methods are left with a difficult task, for which complex inference algorithms have been developed.

Human figures exhibit substantial variety in appearance, particularly due to clothing differences. Textured clothing is quite problematic for methods which use edge features for pose estimation. However, for motion estimation textured clothing is particularly advantageous, as it leads to more reliable motion estimates by reducing aperture effects.

Another advantage to our approach is the use of exemplars to enforce global pose consistency in our tracking algorithm. Our method first finds a sequence of exemplars which match the input sequence. Given these, ambiguities inherent in kinematic tracking from 2d data (such as the left limb - right limb ambiguity) are conveniently dodged. If the sequence of exemplars form a consistent track, the inference of joint positions is left as a simpler task.

This global consistency from exemplars comes at a price however. It is unreasonable to assume that a sufficiently large set of exemplars would exist to enable tracking people performing a variety of actions. However, for limited domains, it is possible to obtain such a sufficient set. In particular, we perform experiments on the CMU MoBo dataset [11], showing the ability of our method to track a variety of people performing simple walking motions. Further, being able to accurately estimate the pose of a person, only in a limited set of poses would be useful for tasks such as initializing a more general kinematic tracker (e.g. [10]).

The main contribution of this paper is developing a self-initializing kinematic tracker based on this motion exemplar framework. We show how we can efficiently perform inference in it with an approximate inference method, first finding a sequence of exemplars and then refining positions of all joints with a Gibbs sampling and gradient ascent scheme.

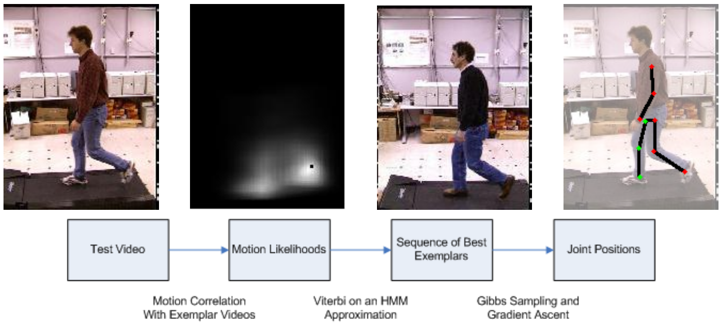
Test Video
Motion Likelihoods
Sequence of Best Exemplars
Joint Positions
Motion Correlation With Exemplar Videos
Viterbi on an HMM Approximation
Gibbs Sampling and Gradient Ascent

Figure 1. Data flow for our algorithm. We compute the motion likelihood for different exemplars at different joint places. The likelihoods are then used to compute the best sequence of exemplars. We use Gibbs sampling and gradient ascent to search for the best positions of joints. Best exemplars are used to prune the search space.

The structure of this paper is as follows. We review previous work in Section 2. We describe our motion exemplar model in Section 3, and provide the details of our approximate inference method in Section 4. We describe our experiments in Section 5 and conclude in Section 6.

2. Previous Work

The problem of tracking humans in videos has been the subject of a vast amount of research in the computer vision community. Forsyth et al. [4] provide a comprehensive survey of approaches to this problem.

A common approach is to assume an initialization of the human pose in the first frame of a sequence is given, after which tracking is performed. An early example of this work is Rohr [12] in which tracking is performed by matching the edges of a projection of a 3d body model to those found in the image.

Other researchers followed a similar approach, using motion estimation rather than comparison of edge maps for a tracking phase. Ju et al. [6] learn a parametric flow model based on a 2d "cardboard person" model. Bregler and Malik [2] use a flow model based on 3d kinematic chain model.

Automatic initialization of such trackers has been explored. The W^4S system of Haritaoglu et al. [5] initializes a simplified cardboard person model using a heuristic background subtraction-based method. Urtasun et al. [22] focus on the learning of motion models for specific activities, and initialize their tracker with simple detectors or by hand. Ramanan et al. [10] initialize with a shape template matcher in order to learn a person-specific appearance model which can be used for tracking.

Our work falls into a category of approaches which simultaneously detect and track. Rosales and Sclaroff [13] describe the Specialized Mappings Architecture (SMA), which incorporates the inverse 3D pose to silhouette mapping for performing inference. Agarwal and Triggs [1] also directly learn to regress 3D body pose. They use shape features extracted from silhouettes, and employ Relevance Vector Machines for regression. Sminchisescu et al. [17] learn a discriminative model which predicts a distribution over body pose from silhouette data, and propagate this distribution over a temporal sequence. Since the silhouette-body pose mapping is ambiguous and multi-modal, complex algorithms for propagating this distribution are required. Sigal et al. [16] and Sudderth et al. [19] track people and hands respectively, using *loose-limbed models*, models consisting of a collection of loosely connected geometric primitives, and use non-parametric belief propagation to perform inference. Sudderth et al. build occlusion reasoning into their hand model. Sigal et al. use *shouters* to focus the attention of the inference procedure.

Another line of approaches infers human pose by matching to a set of stored exemplars by matching using shape cues. Toyama and Blake [21] develop a probabilistic exemplar tracking model, and an algorithm for learning its parameters. Sullivan and Carlsson [20] and Mori and Malik [8] directly address the problem of pose estimation. They stored sets of 2D exemplars upon which joint locations have been marked. Joint locations are transferred to novel images using shape matching. Shakhnarovich et al. [14] address variation in pose and appearance in exemplar matching through brute force, using a variation of locality sensitive hashing for speed to match upper body configurations of standing, front facing people in background subtracted image sequences. Our approach is similar to these methods, but uses motion exemplars rather than shape in order to avoid the aforementioned difficulties due to appearance.

There is a large body of work on matching human motion templates, particularly focused on matching the periodicity present in the human gait. An early example of this work is Niyogi and Adelson [9], who analyze periodic structure of surfaces in XYT volume. While we experiment on walking videos, our approach is not limited to periodic motions, and does not use such assumptions for estimating pose.

Other methods for initializing pose estimates from image sequences include Song et al. [18], who detect corner features in image sequences and model their joint position and velocity statistics using tree-structured models. Dimitrijevic et al. [3] match short sequences of static templates, compared using Chamfer distance.

Our inference method, which first finds a sequence of exemplars to enforce global pose consistency is related to methods such as Lee and Chen [7] for building an interpretation tree for resolving the ambiguity regarding foreshortening (closer endpoint of each link) for the problem of 2d to 3d lifting of joint positions. In our case we find a single most likely sequence of exemplars, but one could reason about other possible sequences instead.

2.1. Motion Correlation

Given a collection of stored exemplar videos, each exemplar sequence is tested to verify how well it matches the input video in some place (x,y,t) in space-time domain. Different people with different clothes and different surrounding background, but in similar poses, can produce completely different space-time intensity patterns in an input video. To solve this problem the method presented in Shechtman and Irani [15] is used to compare the input video by checking the motion consistency between a stored exemplar video with video segments centered around every space-time point. In this section we briefly review this motion consistency measurement, paraphrased from [15].

The consistency between two video segments is evaluated by computing and integrating local motion consistency measures between small space-time patches within the video segments. For each point in each video segment, the motion in space-time patch centered on that point is compared against its corresponding space-time patch in the other segment. The computed local scores are then aggregated to provide a correlation score for the entire segment at that video location.

The motion in every small patch is assumed to be continuous and in a single direction in space-time. To compare the motion consistency between two small patches, they compute their two dimensional ($M_1^\diamond$, $M_2^\diamond$) and three dimensional (M_1, M_2) Gram matrix of gradients in space and space-time. if the direction of motion in two patches are consistent the rank increase of their addition from two dimensional ($M_{12}^\diamond$) to three dimensional (M_{12}) will be close to the minimum of their rank increase. They define $\frac{1}{m_{12}}$ as the consistency score and m_{12} is computed as below,

$$m_{12} = \frac{\Delta r_{12}}{min(\Delta r_1, \Delta r_2) + \epsilon} \tag{1}$$

Where Δr_k is the rank increase from $M_k^\diamond$ to M_k. The minimum value of m_{12} is 1 so $\frac{1}{m_{12}}$ will be always in $[0, 1]$. This helps them to avoid the fundamental hurdles of optical flow estimation (aperture problem, singularities, *etc.*) and makes their method robust to different textures, colors and backgrounds.

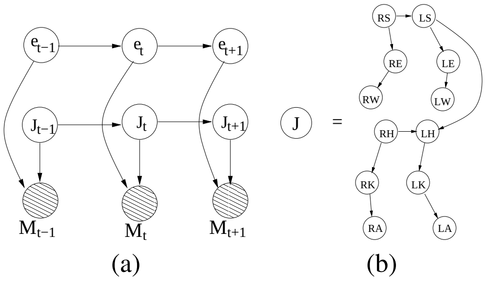

Figure 2. (a) Full graphical model used for inference of joint positions. (b) Each node J consists of body joints with kinematic tree connections within a frame, and temporal connections only between corresponding body joints (not shown).

3. Motion Exemplar Tracking Model

Our algorithm estimates the pose of a human figure in an image sequence by performing motion correlation between the input sequence and the body joints of a set of labeled exemplar sequences. We use a generative model of these motion correlation values, depicted in Figure 2. Using exemplars will remove the ambiguities inherent in kinematic tracking from 2d data. In this section we provide the details of this model.

We will use the following notation in this description. e_t will denote the exemplar used at time t. For clarity of presentation, J_t will be used to denote the set of 12 2d body joint positions at time t, which are connected in a kinematic structure shown in Figure 2(b). M_t is the set of all exemplar-input frame motion correlation measurements at time t. Again, there is structure to these measurements which is not depicted in Figure 2 for clarity, but which will be described below.

3.1. Motion Correlation Likelihood

In this section we describe our model for the likelihood of observing a particular set M_t of motion correlation measurements given an exemplar e_t and set of joint positions J_t. We perform correlation using space-time windows centered around each body joint in each exemplar, and others at larger scales. We formulate a likelihood model in which each joint generates the motion correlations in its position.

Dropping the subscript t for clarity, let $M = m_i^{k,s}$ be the set of exemplar joint-pixel motion correlations. $m_i^{k,s}$ is the correlation between window s on exemplar k with the input image at pixel i. In our experiments, index s runs over 3 scales of windows for each of the 12 body joints.

We make the usual independence assumption to model the likelihood $P(M|J, e)$. We assume the elements of M to be conditionally independent given J and e. For us, this assumption is reasonable, as the larger scale correlations from the exemplar gives global structure to the motion responses:

$$P(M|J,e) \quad = \prod_{(i,k,s)\in(\mathcal{P},\mathcal{E},\mathcal{S})} P(m_i^{k,s}|J,e) \qquad (2)$$

where $\mathcal{P}$ is the set of pixel indices in an image, $\mathcal{E}$ is the set of exemplar indices, and $\mathcal{S}$ is the set of correlation windows.

We split this set of motion correlations into a foreground set $\mathcal{F} = \{(i,k,s)\}$, containing all pixels-windows (i,s) corresponding to a body joint in J, with exemplar $k = e$, and background set $\mathcal{B}$, containing the remainder:

$$\begin{aligned} P(M|J,e) &= \prod_{(i,k,s)\in\mathcal{F}} P_{fg}(m_i^{k,s}|J,e) \prod_{(i,k,s)\in\mathcal{B}} P_{bg}(m_i^{k,s}|J,e) \quad (3) \\ &\propto \prod_{(i,k,s)\in\mathcal{F}} \frac{P_{fg}(m_i^{k,s}|J,e)}{P_{bg}(m_i^{k,s}|J,e)} \quad (4) \end{aligned}$$

We will model these two distributions using separate Gaussians, the foreground distribution P_{fg} for motion correlations corresponding to the proposed body joint location and exemplar, versus the P_{bg} for those corresponding to background locations.

The parameters of these distributions are fit with training data. For each joint of each exemplar in the training set, we find the highest correlation value with an exemplar from another person in the training set. This set of correlations becomes our positive training set, and we fit a Gaussian to these values. For the background distribution we randomly sample a set of non-matching correlation values.

3.2. Exemplar Transition Model

The probability of transition from an exemplar $e_{t-1} = h$ to another exemplar $e_t = k$ is computed by comparison of the angles and also angular velocities of their limbs. We use angles rather than joint positions to be able to compare exemplars from different people while ignoring their variation in size. For each limb j in each exemplar k, a 2d angle $\theta^j(k)$ and its angular velocity $\dot{\theta}^j(k)$ are computed, the latter by examining the preceding frame. To find the transition probability $P(e_t = k|e_{t-1} = h)$, the angular change and the angular velocity change of the limbs are assumed to follow a Gaussian distribution.

$$\begin{aligned} &P(e_t = k|e_{t-1} = h) \\ &\propto \prod_{j\in\mathcal{L}} e^{-[(\theta^j(k)-\theta^j(h))-\mu_j]^2/2\sigma_j^2} e^{-[(\dot{\theta}^j(k)-\dot{\theta}^j(h))-\dot{\mu}_j]^2/2\dot{\sigma}_j^2} \quad (5) \end{aligned}$$

Where $\mathcal{L}$ is the set of all limbs in an exemplar. The parameters of the Gaussian distribution are fit using training data. For all exemplars which come from adjacent frames k and h we calculate $\theta^j(k) - \theta^j(h)$ and $\dot{\theta}^j(k) - \dot{\theta}^j(h)$. These sets of values become our positive training data and a Gaussian is fit to each one. Note that e_t, the exemplar used in a particular frame, is not grounded at any particular location, and hence, relationships between body joints, spatially and temporally, must be modeled, which will be described next.

3.3. Dependencies Between Body Joints

J_t consists of a set of 12 body joint positions: shoulders, elbows, wrists, hips, knees, feet. Every joint position J_t^j in frame t is connected to its corresponding joint position J_{t-1}^j in frame $t-1$. In addition, it is connected to some other joint position in frame t under the kinematic tree in Figure 2(b). The motion model is computed by using a two dimensional Gaussian. For the spatial prior between joint J_t^j and its parent $J_t^{\pi(j)}$, a simple uniform distribution over a disk is used to enforce connectivity.

$$\begin{aligned} P(J_t^j|J_{t-1}^j, J_t^{\pi(j)}) = \quad & \mathcal{N}(J_t^j - J_{t-1}^j; \mu, \Sigma) \cdot \\ & \mathcal{U}(J_t^j - J_t^{\pi(j)}; r_{min}, r_{max}) (6) \end{aligned}$$

The parameters of the two dimensional Gaussian distribution for each joint type are set by the mean and covariance of its displacement in adjacent frames of training data. The maximum and minimum radius of the disk are set to the maximum and minimum distance found from the training data.

4. Inference

Exact inference in the model we described in the previous section is not tractable. The temporal connections between body joints J_t and the dependence between body joints and exemplars e_t would lead to a loopy graph for message passing algorithms. In addition, since the image likelihoods are multi-modal, straight-forward techniques such as Kalman filters would not be applicable. Instead, we use an approximate inference procedure.

In the following sections we describe this procedure. We first fix the exemplars to be used in each frame by finding the best sequence of exemplars using the Viterbi algorithm, on an approximation of our model. Fixing the exemplars will help us to reduce the huge search space into a manageable one. In addition, it will give us a good initial estimate for the positions of the individual joints. From this initial estimate, we then perform a sampling procedure to obtain a set of samples of possible body joint configurations. The uncertainty that needs to be captured by this sampling procedure is less than the original inference procedure since we have restricted ourselves to a particular sequence of exemplars. Modes from this sampled distribution are then found, and each is locally optimized using gradient ascent.

4.1. Exemplar Sequence Estimation

The first step in our approximate inference method is to find a sequence of exemplars. We desire to find the best sequence of exemplars given the observed motion correlations:

$$\hat{e}_{1:t} = \arg\max_{e_{1:t}} P(e_{1:t}|m_{1:t}) \quad (7)$$

$$= \arg\max_{e_{1:t}} \int_{J_{1:t}} P(e_{1:t}, J_{1:t}|m_{1:t}) \quad (8)$$

where $e_{1:t}$ denotes the sequence from time 1 to time t.

However, performing the above integral over sequences $J_{1:t}$ is not practical. Instead, we make two simplifying assumptions from our model in order to compute this sequence. These assumptions are made with the intent of converting our model to be similar to a simple Hidden Markov Model (HMM) for which sequence estimation is straightforward.

The first assumption is to select, for each frame, a set $\hat{J}_t^k$ for each value of the exemplar random variable e_t which maximizes the likelihood $P(M_t|\hat{J}_t^k, e_t = k)$. Since our model for the body joints in each frame is tree-structured, this inference task is computationally efficient. Now, instead of considering all possible sets of locations for all body joints, for a particular exemplar we will limit ourselves to this one set.

The second is to ignore the temporal connections in our model at the level of joints. Temporal connections at the exemplar level will still be included, and therefore overall global consistency of the track will be maintained.

The integral in Equation 8 is then approximated as

$$\int_{J_{1:t}} P(e_{1:t}, J_{1:t}|m_{1:t}) \approx P(e_{1:t}, \hat{J}_{1:t}|m_{1:t}) \quad (9)$$

$$\approx \prod_{i=1}^{t} P(m_i|e_i, \hat{J}_i^{e_i})P(e_i|e_{i-1}) \quad (10)$$

Finding the most likely sequence of this product is then a straight-forward dynamic programming problem, akin to HMM decoding using the Viterbi algorithm.

4.2. Gibbs Sampling / Gradient Ascent

The previous exemplar sequence inference procedure will result in a sequence of $\{\hat{e}_t, \hat{J}_t\}$ values. These body joint position sequences are smooth at the level of angles because the exemplars are consistent. However, they might not match the person in this frame accurately since actual joint positions have not yet been taken into account. The next step is to perform inference of J_t using the temporal model $P(J_t|J_{t-1})$ which was omitted from the previous step.

Even after fixing the sequence of exemplars, exact inference is still intractable due to the temporal links between body joint positions. Instead, we employ Gibbs sampling, a Markov Chain Monte Carlo algorithm, to obtain samples from the distribution $P(J_{1:t}|M_{1:t}, \hat{e}_{1:t})$.

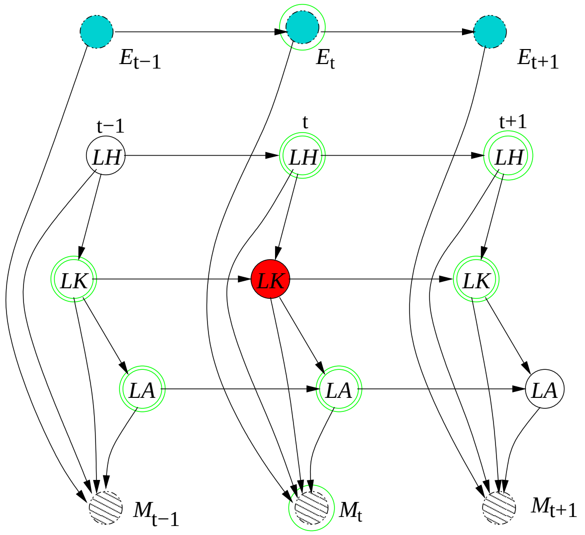

Figure 3. Gibbs sampling procedure. A simplified three node kinematic chain is shown for clarity. Node LK_t, shown in red, has been chosen for sampling at this iteration. Exemplars, in blue, have been fixed via the approximation scheme. All other nodes have current values, a new value for LK_t is chosen by sampling from the marginal over LK_t. Only nodes in its Markov blanket, shown with green circles, need to be considered.

We initialize the state of our model to the $\{\hat{e}_t, \hat{J}_t\}$ sequence. At each step of the Gibbs sampling we choose a particular joint J_t^i to change, and set its value by sampling from the conditional distribution $P(J_t^i|\bar{J}, M_{1:t}, \hat{e}_{1:t})$, where $\bar{J}$ denotes all joints other than J_t^i. The mentioned conditional distribution is computed by multiplying all conditionals involving the Markov blanket of J_t^i. Each of these conditionals is essentially a 1-D function, as all other joints are fixed. Figure 3 illustrates the computation of this conditional distribution.

This Gibbs sampling procedure is employed to handle the remaining ambiguity, although many of the disparate modes are already eliminated by the exemplar inference procedure. Sampling is not guarantied to find the global maxima, as a result we run this Gibbs sampling procedure, collect the modes of the samples, and for every mode run gradient ascent step to produce an estimate of the best sequence of joint positions $J_{1:t}$. The sequence with the highest posterior is then returned as our result.

5. Results

Experiments are performed on different subsets of images from the CMU Mobo database [11]. We have tested our algorithm on four different sequences: side-view fast walk (fastWalk/vr03_7), 45^o-view fast walk (fastWalk/vr16_7), side-view incline walk (incline/vr03_7) and side-view slow walk (slowWalk/vr03_7). 9 subjects (numbers 04006-04071), 30 frames each, are selected from the aforementioned sequences. Marking of exemplar joint locations was performed manually for all four collections of 270 frames. This dataset enables us to study the robustness of our method handle variations in body shape, clothing, and viewpoint.

For each sequence, a set of 9 experiments was performed in which each subject was used once as the query against a set of different exemplar videos extracted from remaining eight subjects (leave-one-out cross validation). Three scales of space-time windows are used for each joint, of sizes similar to the whole body, lower or upper body limbs, and around each joint are used to create the motion correlation data. Each space-time window is 3 frames long. As each subject consists of 30 frames, there will be 8×28 windows of each kind (8 exemplars). These exemplar videos are correlated with query video in all possible positions in space and time to compute the motion likelihood. As space-time correlation is computationally expensive we have used coarse to fine search to enhance the speed. We have used $7 \times 7 \times 3$ small patches around each pixel. The size of the small patches represents the cells that can be assumed to have a continuous motion.

We found experimentally that it was advantageous to use separate exemplars for upper and lower body joint positions, rather than a single exemplar for the entire body. Splitting the exemplars in this fashion helps by reducing the amount of variation a single exemplar needs to capture. The inference procedure is described above, except two separate runs of the Viterbi algorithm are used. It would also be possible to perform this inference of two exemplars per frame jointly.

The best sequence of exemplars for the upper body limbs (left and right arm) and lower body limbs (left and right legs) are found by applying Viterbi algorithm. Fig. 4 shows sample results of best sequence of upper and lower exemplars for different sequences. The sequence of best exemplars works well to discriminate between left and right limbs especially for side-view sequences for which limb labels can be ambiguous.

Having the exemplars fixed, joint positions are initialized by maximizing the likelihood at every single frame ignoring the connections between adjacent frames. We start moving from frame 1 to 30 and in each frame from top to down and using Gibbs sampling to sample each node. We perform this iteration 60 times and every time the result is fed to gradient ascent to find the local maxima. Finally the sampled configuration that maximizes the likelihood of whole graph is chosen as the result.

Left limb joint positions are shown by red dots and right limb joints with green. Note that on the side view sequences the right arm is mostly occluded and is therefore not included in our model. Example results are shown in Fig. 5.

Our results for side view fast walk are compared to Mori and Malik [8] in Table 1. Our method significantly outperforms their shape context matching exemplar method. As our method is based on motion correlation it is more precise for end limbs such as elbows, wrists, knees and feet where there is always movement rather than shoulder, hip and head. Our method significantly outperforms shape context matching for subjects who wear loose-fitted trousers (which produce irregular folds) or have a textured shirt, such as subjects 4006, 4022 and 4070 (rows 1, 5, and 8 in Table 1).

In the subset of CMU MoBo used in Mori and Malik [8], upon which these quantitative results are based, different subjects usually have similar patterns of movement in their legs, but arm movement can be quite irregular. People have different velocities, configurations and various amount of rotation in their arms (*e.g.* the variation of angle between arm and body is substantial in 4011 and 4022). These irregularities lead to less accurate joint position estimation for the arms than for the legs, though a larger training set would likely alleviate these problems.

6. Conclusion

In this paper we have presented a novel motion-exemplar framework for building a self-initializing kinematic tracker for human figures. The use of motion estimates has advantages over previous methods which use edge comparisons for image likelihoods, and those which use silhouette features. We presented quantitative results demonstrating this, showing that our method outperforms an exemplar method using shape matching, particularly for human figures wearing textured clothing.

We believe that promising future directions for research include combining exemplars from multiple viewpoints, and experimenting with the sensitivity of our method to viewpoint variation between the exemplars and the input video. Further, we believe that this exemplar-based tracking could be used to initialize a more general tracker, either one using more precise motion models (e.g. [22]) or person-specific appearance models (e.g. [10]).

References

[1] A. Agarwal and B. Triggs. 3d human pose from silhouettes by relevance vector regression. In *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.*, 2004. 1, 2

[2] C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.*, pages 8–15, 1998. 2

[3] M. Dimitrijevic, V. Lepetit, and P. Fua. Human body pose detection using bayesian spatio-temporal templates. *Computer*

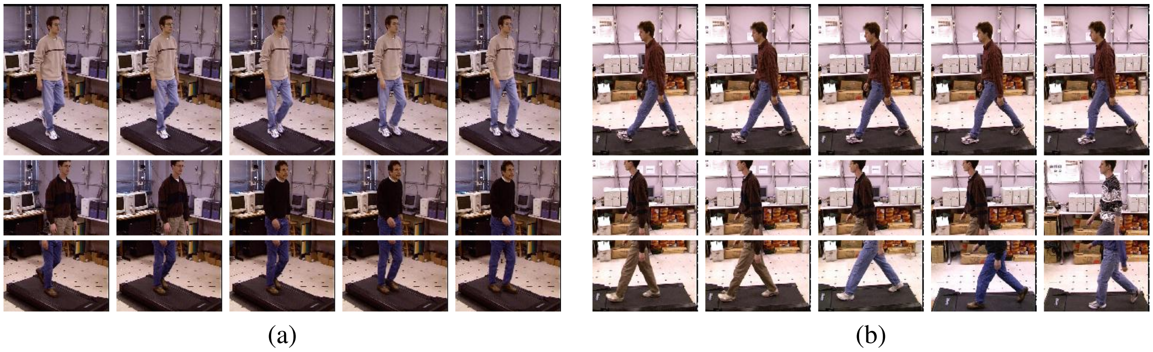
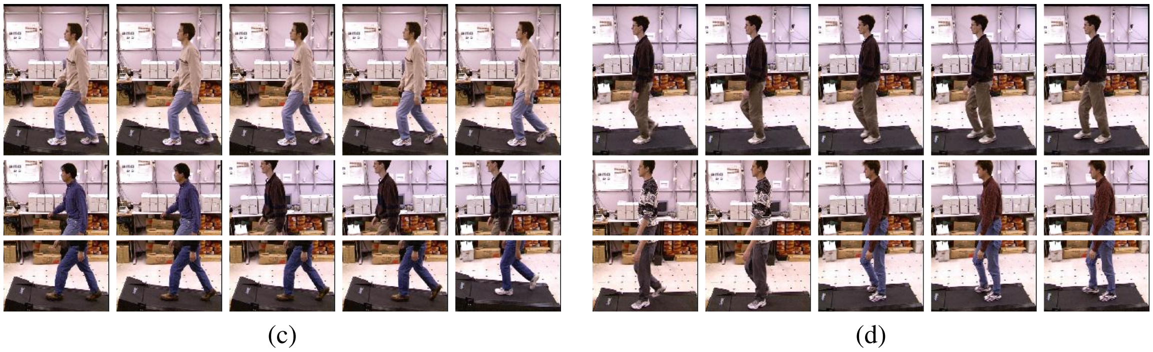

Figure 4. The Viterbi algorithm is used to find the best sequence of upper and lower exemplars: (a) 45^o-view fast walk, (b) side-view fast walk, (c) side-view incline, (d) side-view slow walk.

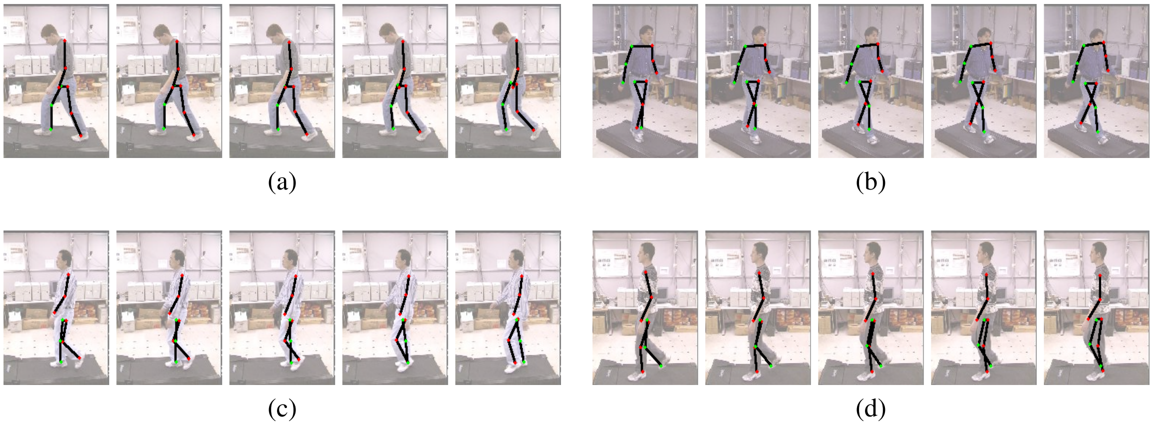

Figure 5. Sample results for finding the body joints. Left limb joints are shown with red dots and right limb joints are presented in green. (a) side-view incline, (b) 45^o-view fast walk, (c) side-view fast walk, (d) side-view slow walk.

Vision and Image Understanding, 104(2):127–139, December 2006. 3

[4] D. A. Forsyth, O. Arikan, L. Ikemoto, J. O'Brien, and D. Ramanan. Computational studies of human motion: Part 1, tracking and motion synthesis. *Foundations and Trends in Computer Graphics and Vision*, 1(2), 2006. 2

[5] I. Haritaoglu, D. Harwood, and L. Davis. W4s: A real time system for detecting and tracking people in 2.5d. In *European Conference on Computer Vision*, 1998. 2

[6] S. Ju, M. Black, and Y. Yacoob. Cardboard people: a parameterized model of articulated image motion. In *IEEE International Conference on Face and Gesture Recognition*, pages 38–44, 1996. 2

[7] H. J. Lee and Z. Chen. Determination of 3d human body posture from a single view. *Comp. Vision, Graphics, Image Processing*, 30:148–168, 1985. 3

[8] G. Mori and J. Malik. Recovering 3d human body configurations using shape context matching. *IEEE Trans. PAMI*, 28(7):1052–1062, 2006. 1, 2, 6, 8

[9] S. A. Niyogi and E. H. Adelson. Analyzing gait with spa-

	Shoulder	Elbow	Hand	Hip	Knee	Ankle
	10.4 ± 7 12.9 ± 9	16.4 ± 9 27.0 ± 26	14.4 ± 8 43.3 ± 47	20.9 ± 10 18.1 ± 11	16.1 ± 6 32.9 ± 36	14.2 ± 9 45.8 ± 59
	25.6 ± 7 15.1 ± 7	29.0 ± 13 18.4 ± 9	45.6 ± 17 26.5 ± 15	12.3 ± 7 16.6 ± 7	16.1 ± 8 16.4 ± 10	17.6 ± 14 16.4 ± 12
	25.3 ± 6 12.9 ± 13	33.0 ± 10 22.3 ± 20	34.0 ± 19 27.8 ± 20	17.2 ± 6 15.2 ± 7	10.9 ± 6 13.6 ± 7	15.8 ± 11 17.4 ± 22
	8.7 ± 5 13.8 ± 8	20.6 ± 13 29.4 ± 16	22.4 ± 10 27.4 ± 25	11.2 ± 8 22.8 ± 14	11.5 ± 5 24.4 ± 25	10.9 ± 7 24.1 ± 23
	21.2 ± 9 42.8 ± 24	30.2 ± 17 69.1 ± 40	52.8 ± 27 98.7 ± 54	13.4 ± 6 41.8 ± 23	11.2 ± 7 56.1 ± 27	12.6 ± 8 79.0 ± 55
	21.8 ± 8 13.8 ± 9	14.4 ± 7 31.9 ± 17	36.8 ± 25 34.7 ± 30	18.4 ± 11 16.1 ± 8	12.6 ± 6 19.6 ± 20	12.1 ± 7 26.8 ± 42
	9.2 ± 6 13.2 ± 8	18.1 ± 8 24.0 ± 10	23.2 ± 10 27.7 ± 37	10.2 ± 5 15.6 ± 9	12.3 ± 5 21.0 ± 28	12.7 ± 5 24.7 ± 48
	18.3 ± 8 17.9 ± 11	24.0 ± 12 34.4 ± 33	20.3 ± 9 45.8 ± 41	11.1 ± 5 25.7 ± 12	13.3 ± 9 34.6 ± 29	15.1 ± 7 45.9 ± 50
	17.3 ± 7 17.9 ± 8	22.6 ± 11 22.6 ± 13	20.3 ± 10 28.5 ± 24	21.0 ± 7 16.0 ± 9	15.1 ± 6 20.7 ± 20	15.0 ± 5 23.5 ± 37
Mean **Mean**	17.5 17.8	23.1 31.0	30.0 40.0	15.1 20.9	13.2 26.5	15.0 33.7

Table 1. Mobo database ”fast walk side-view” subject numbers versus joint position error. The error is based on the distance of computed position from ground truth in pixels. Each cell shows error mean and standard deviation by our proposed method (top) and that of shape context exemplars [8] (bottom). The last row shows the mean error over all 9 subjects.

tiotemporal surfaces. In *IEEE Workshop on Nonrigid and Articulated Motion*, 1994. 2

[10] D. Ramanan, D. Forsyth, and A. Zisserman. Strike a pose: Tracking people by finding stylized poses. In *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.*, 2005. 1, 2, 6

[11] R.Gross and J. Shi. The CMU motion of body (MoBo) database. Technical Report CMU-RI-TR-01-18, Robotics Institute, Carnegie Mellon University, 2001. 1, 5

[12] K. Rohr. Incremental recognition of pedestrians from image sequences. In *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.*, pages 8–13, 1993. 1, 2

[13] R. Rosales and S. Sclaroff. Learning body pose via specialized maps. In *Neural Information Processing Systems NIPS-14*, 2002. 1, 2

[14] G. Shakhnarovich, P. Viola, and T. Darrell. Fast pose estimation with parameter sensitive hashing. In *Proc. 9th Int. Conf. Computer Vision*, volume 2, pages 750–757, 2003. 1, 2

[15] E. Shechtman and M. Irani. Space-time behavior based correlation. In *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.*, 2005. 3

[16] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isard. Tracking loose-limbed people. In *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.*, 2004. 2

[17] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas. Discriminative density propagation for 3d human motion estimation. In *Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn.*, 2005. 1, 2

[18] Y. Song, L. Goncalves, and P. Perona. Unsupervised learning of human motion. *IEEE Trans. PAMI*, 25(7):814–827, 2003. 2

[19] E. Sudderth, M. Mandel, W. Freeman, and A. Willsky. Distributed occlusion reasoning for tracking with nonparametric belief propagation. In *NIPS*, 2004. 2

[20] J. Sullivan and S. Carlsson. Recognizing and tracking human action. In *European Conference on Computer Vision LNCS 2352*, volume 1, pages 629–644, 2002. 1, 2

[21] K. Toyama and A. Blake. Probabilistic exemplar-based tracking in a metric space. In *Proc. 8th Int. Conf. Computer Vision*, volume 2, pages 50–57, 2001. 1, 2

[22] R. Urtasun, D. Fleet, and P. Fua. Temporal motion models for monocular and multiview 3d human body tracking. *Computer Vision and Image Understanding*, 104(2-3):157–177, 2006. 2, 6