A Unified Spatio-Temporal Articulated Model for Tracking

Xiangyang Lan Daniel P. Huttenlocher
{xylan,dph}@cs.cornell.edu
Cornell University
Ithaca NY 14853

Abstract

Tracking articulated objects in image sequences remains a challenging problem, particularly in terms of the ability to localize the individual parts of an object given self-occlusions and changes in viewpoint. In this paper we propose a two-dimensional spatio-temporal modeling approach that handles both self-occlusions and changes in viewpoint. We use a Bayesian framework to combine pictorial structure spatial models with hidden Markov temporal models. Inference for these combined models can be performed using dynamic programming and sampling methods. We demonstrate the approach for the problem of tracking a walking person, using silhouette data taken from a single camera viewpoint. Walking provides both strong spatial (kinematic) and temporal (dynamic) constraints, enabling the method to track limb positions in spite of simultaneous self-occlusion and viewpoint change.

1. Introduction

We consider the problem of detecting and tracking an articulated object, such as a person in a video sequence, using a two-dimensional parameterized model. The main focus of our work is on handling self-occlusions and changes in viewpoint, which remain challenging problems in the tracking of articulated objects. Our approach combines pictorial structure spatial models and hidden Markov temporal models into a single unified framework. A pictorial structure [5] consists of a set of rigid parts where certain parts are connected by springs to allow for bending and stretching. The parts and their constraints are two-dimensional, similar to the cardboard people models used in [7] for tracking people, but within a more general spatial modeling framework. Hidden Markov models (HMM's) have been used extensively for tracking as well as for gait and gesture recognition. In our approach the states of the HMM correspond to parameterized pictorial structure models that capture spatial characteristics of representative views or "key frames". The transitions between states capture temporal dynamics based on interpolation between representative views.

The main distinguishing characteristic of our work is the integration of parameterized spatial models and temporal models into a common framework. When the states of an HMM are themselves parameterized models, it is generally intractable to perform inference – e.g., to find the best sequence of states for a given input sequence, or even to find good such state sequences. Here we provide a good inference procedure for object detection and tracking, when the states of the HMM are pictorial structure spatial models. The resulting combined model provides both strong spatial (or kinematic) constraints from the pictorial structures and strong dynamic constraints from the HMM. The model clearly delineates between the spatial and dynamic constraints, yet the two are brought together in a common estimation framework.

There are a wide range of approaches to detection and tracking, many of which are surveyed in [6]. In contrast with our approach, most prior work that uses parameterized-state HMM's for tracking articulated objects is based on having linear dynamic system (LDS) models for the states (e.g., [1, 8]). Such methods have strong dynamic constraints but relatively weak spatial constraints, because the LDS models (and associated Kalman filtering techniques) capture primarily dynamic rather than spatial characteristics. HMM's have also been used in conjunction with template spatial models, for problems in gesture, gait and action recognition (e.g., [3]) These methods differ from our work in that each state is a template rather than a parameterized model, making it difficult to adapt such methods to tracking the actual configuration of parts for an articulated object. Parametric HMM's have been used in other contexts, such as gesture recognition (e.g., [12]). But their work is more focused on model learning, and difficult to apply to high-dimensional problems such as tracking articulated objects.

Pictorial structure models have recently been successfully applied to tracking humans [10], but without explicit models of the temporal constraints. Our work extends those results by using an HMM temporal model, and as a consequence is able to handle changes in viewpoint and self-

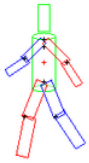

Figure 1: *The default configuration for a simple pictorial structure model corresponding to a side view of a person walking. Red denotes the person's left limbs, blue their right limbs and green their head and torso.*

occlusions. The focus in [10] is on learning good appearance models for the parts of an object, for use in tracking, whereas our focus is on combining spatial and temporal constraints. Thus the two pieces of work are complementary. Another recent approach to tracking articulated objects uses a tree-based filtering method [11]. They demonstrate the method for hand tracking but it could also be applied to other articulated objects. In their approach a large number of views are clustered together and the focus of the work is on using a tree structure to rapidly search for the best matching view. In contrast we use a small number of representative views for accurate tracking of multiple parts, by having a highly parameterized pictorial structure model.

2. An Integrated Spatio-Temporal Articulated Model

We now discuss the specifics of the pictorial structure and HMM models, and how they fit together in an integrated modeling framework. A pictorial structure is a parametric spatial model composed of multiple parts, with spring-like connections between certain pairs of parts. An instance of such a model consists of the parts $P = (p_1, \ldots, p_n)$ and the connections $C = \{c_{ij}\}$, where each c_{ij} represents the spatial constraints between parts p_i and p_j. Each part p_i has parameters representing the appearance of that part. For instance the appearance of a part may be characterized by local oriented filter responses, template models, feature detectors, etc. Each connection c_{ij} has parameters that encode both the ideal relative spatial configuration of two parts and spring-like constants controlling deformations from this ideal configuration. For instance the spatial relations may specify a relative position and orientation together with covariances that characterize the degree of bending and stretching between the parts.

If we let $\mu = (P, C)$ denote such a pictorial structure model, then a *configuration* of the model is given by $\Theta = (\theta_1, \ldots, \theta_n)$ where the θ_i are parameters specifying the location of each part p_i in an image coordinate frame. In a Bayesian framework the prior distribution of configurations for a given model, $P(\Theta|\mu)$, characterizes the probability of different configurations occurring, independent of any observed data. This prior is governed by the connection parameters C, as described in [4]. The highest probability such configuration can be thought of as the preferred or "default" configuration of the model. Such a configuration is illustrated in Figure 1 for a model corresponding to a side view of a person walking, when their arms and legs are maximally extended. We use the color red to denote the left arm and leg, blue to denote the right arm and leg, and green to denote the torso and head. Deviations from this maximal prior probability configuration can be thought of as incurring a cost, in that such configurations have a lower chance of occurring.

The posterior distribution of the configuration parameters,

$$P(\Theta|J, \mu) \ ,$$

characterizes the probability of different configurations given a particular model μ and image J. There are several standard definitions of what constitute good configurations. One definition is the *MAP estimate*, which is a set of parameter values that maximize the posterior probability. Another definition is any set of parameter values for which the posterior probability is high. Such values are generally computed using *sampling* methods, where configurations are selected at random, weighted by their probability. Generally some other technique is used to select the best of these sampled configurations. Following [4] we use the Chamfer distance as a selection criterion.

Using Bayes' rule, the posterior can be expressed in terms of the product of a likelihood and prior,

$$P(J|\Theta, \mu)P(\Theta|\mu) \ . \tag{1}$$

In general for a high-dimensional parameter space Θ it is not tractable to find the Θ^* that maximizes this probability, or to sample values of Θ that have high probability. In the case of pictorial structure models where the parts have a tree-like skeletal structure, such as a human body or hand, both these problems can be solved efficiently using dynamic programming methods as described in [4].

We now turn from the spatial to the temporal component of the models. Let $S = \{s_1, \ldots, s_n\}$ be a set of representative views of an object. Representative views are analogous to "key frames" in animation, in that a video sequence can be summarized by a sequence of representative views. Intermediate frames between two representative views can be generated by interpolation of those views. As noted in the introduction, while representative view temporal models have been used in other work, in these models the views are generally templates or exemplars and are thus not well suited to tracking the configuration of parts of articulated objects. In our case each representative view is a parameterized pictorial structure model rather than a template image.

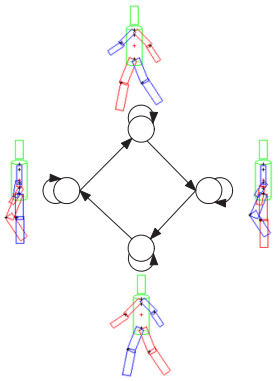

Figure 2: *Four state model corresponding to a side view of a person walking.*

A hidden Markov model (HMM) is a stochastic finite automaton, where each state generates (or accepts) some observation. Let Q_t represent the hidden state and Y_t represent the observation at time t, $1 \leq t \leq T$, where the states take on values in some set $\mathcal{S}$ and the observations take on values in some set $\mathcal{O}$. An HMM is characterized by the tuple $\lambda = (A, B, \pi)$ where the *transition probabilities* are $A(i,j) = P(Q_t = j | Q_{t-1} = i)$ for $i, j \in \mathcal{S}$, the *observation probabilities* are $B(i,k) = P(Y_t = k | Q_t = i)$ for $k \in \mathcal{O}$ and $i \in \mathcal{S}$, and the *initial probabilities* are $\pi(i) = P(Q_1 = i)$ for $i \in \mathcal{S}$. For a good tutorial on HMM's see [9]. In our case the observation is the image at each time t, and each state is a pictorial structure model corresponding to some representative view.

Given an observed sequence of images $Y = (Y_1, \ldots, Y_T)$ and an HMM $\lambda = (A, B, \Pi)$ we want to determine a corresponding best sequence of representative views $Q^* = (Q_1^*, \ldots, Q_T^*)$. This temporal estimation problem is commonly expressed as that of finding a maximum posterior probability (MAP) state sequence

$$Q^* = \text{argmax}_Q P(Q|Y, \lambda) \ ,$$

and can be solved in time $O(Tn^2)$ for a given model λ and observation sequence Y using the Viterbi algorithm.

It is common to consider an HMM in terms of its *state space* graph, where the nodes correspond to states and the arcs correspond to non-zero entries in $A(i,j)$. By way of illustration, Figure 2 shows such a graph for the case of a side view of a person walking (we focus more in depth on walking in Section 3). The Figure shows each state together with the highest prior probability configuration of the corresponding pictorial structure model for that state. The pictorial structures for all four states of this model consist of the same set of parts, P. Only the spatial connection parameters, C, differ between states, as the only differences in the views are in the expected part locations. The states shown at the top and bottom of the Figure correspond to maximal extension of the arms and legs, and differ only in whether the right arm and left leg are forward or vice versa. The states shown at the sides of the Figure correspond to minimal extension of the arms and legs, and differ only in whether the right leg is down and the left leg raised or vice versa.

Note that in this model the two states corresponding to minimal extension of the limbs cannot be distinguished on the basis of a single observed image, because they differ only in which leg is bent. Similarly for the two states corresponding to maximal extension of the limbs. Over time (and given a starting state) these ambiguities can be resolved based on the allowable state transitions indicated by the arrows, thus making it possible to distinguish between states and thereby determine which parts are occluded. This illustrates how the combination of simple parameterized spatial models and temporal models can be used to eliminate ambiguity and localize parts that are occluded by other parts.

2.1. Combining the Spatial and Temporal Models

When the states of an HMM are parameterized models, it is often not tractable to determine a best state sequence, Q^*, because the observation probabilities cannot be reasonably estimated. For HMM's with LDS state models, approximate inference procedures have been developed (e.g, [8]). Here we consider the case of HMM's with pictorial structure state models, where it turns out there is a simple inference procedure. In this combined spatio-temporal model, the observation probability, $B(i,k) = P(Y_t = k | Q_t = i)$, results from the parameterized pictorial structure spatial model, μ_i. This is simply the probability over all possible configurations, which can be determined by integrating over the model parameters,

$$P(Y_t = k|\mu_i) = \int P(Y_t = k|\Theta_i, \mu_i) P(\Theta_i|\mu_i) d\Theta_i \ . \quad (2)$$

Computing this integral when Θ_i is a high-dimensional parameter vector, as is the case here, is generally intractable. A standard approach to approximating such integrals is to sample from the distribution. In our case this sampling can be done efficiently because the distribution is the posterior over configuration parameters in equation (1), which as noted above can be sampled efficiently for a tree-structured model. Moreover, as a byproduct of the sampling methods in [4] the integral in (2) can actually be computed directly without needing to sample configurations.

Given an image sequence $Y = (Y_1, \ldots, Y_T)$, the best state sequence $Q^* = (Q_1^*, \ldots, Q_n^*)$ can thus be computed using the Viterbi algorithm, once the observation probabilities $B(i,k) = P(Y_t = k | Q_t = i)$ for each image and state have been computed using the corresponding distribution of the form in (1). For the best state Q_t^* at each time t, sampling can be used to find parameters Θ^* that correspond to a high probability configuration of the model. Thus the best

state sequence naturally specifies a corresponding sequence of configuration parameter values $\Theta_1^*, \ldots, \Theta_T^*$, specifying a high probability configuration of the best matching model μ_i at each time t. These configurations take into account both the spatial constraints in matching individual pictorial structure models to images and temporal constraints in selecting the best state sequence Q^*.

The above sequence of configuration parameters is based on matching the best representative view model at each time frame. When parts are occluded, interpolation between successive representative views can be used to predict better part locations than are obtained by simply using one of the representative view models. This can be handled easily in the current framework in one of several ways. One approach is to simply introduce additional states into the HMM that are intermediate between representative views. These views have connection parameters, C that are based on interpolation between adjacent representative views. The main disadvantage of this approach is that it increases the number of states and pictorial structure models, which slows down processing.

An alternative approach to obtaining better estimates of the locations of occluded parts is to re-estimate the parameters Θ_t at each time, after first finding the best state sequence without the addition of any interpolated states to the HMM. The observation probability $P(Y_t = k|Q_t^* = i)$ reflects how well the best model μ_i accounts for the observed image at time t. This probability will be lower for time frames that lie between two representative views compared with those that are a good match to a single representative view. This can be used to determine the degree to which an image is intermediate between representative views and then generate a corresponding interpolated model. The configuration can then be estimated by matching this interpolated model to the image rather than the model for the best state. We use this approach to obtain better estimates for the locations of occluded parts.

As described above the transition probabilities A are stationary, and do not vary with time. In our context it could further be useful to have transition probabilities that vary with time. For instance, one could envision stronger dynamic models where the transition probabilities vary. Given such additional estimates the same Viterbi method can be applied as above except $A = (A_1, \ldots, A_t)$ is varying with time.

A drawback of the Viterbi approach is that it operates on an entire sequence of frames at once, whereas in interactive tracking applications it is desirable to do incremental processing one frame at time rather than batch processing on an entire sequence. There are several standard ways of handling this including using the forward probabilities from the forward-backward algorithm, rather than the Viterbi method. This corresponds to summing rather than maximizing products of probabilities.

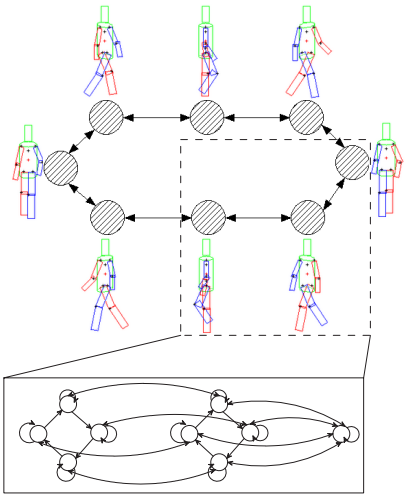

Figure 3: *Modeling transitions between viewpoints.*

3. Application to Human Walking

In the remainder of the paper we consider the problem of tracking a walking person, including determining the locations of their limbs, using silhouette data taken from a single camera viewpoint. Human walking has relatively simple temporal characteristics, yet this problem is still challenging due to single-frame ambiguities, self-occlusion and changes in viewpoint. This section will show how the pictorial structure based HMM framework introduced above can be used to tackle these difficulties effectively.

3.1. A Model of Walking

When only a single viewpoint is considered the simple four state model introduced above, and shown for a side view in Figure 2, provides a generic description of walking as a cyclical gait between four representative views. Similar four-state models have been proposed and used in a wide variety of contexts, both in computer vision and in studies of human walking.

To allow for changes in viewpoint, we consider the eight possible viewing directions of front, back, left, right and 45-degree views halfway between them (front-left, front-right, back-left and back-right). The only allowable state transitions between viewpoints are to stay at the same viewpoint or to change to one of the immediately adjacent viewpoints. These states and transitions are illustrated in the graph at the top of Figure 3. Each node in this graph actually corresponds to an entire cyclical gait model. We use the four-state gait model in Figure 2 for the two side views, a similar model for the four 45-degree views, and a one-state model for the front and rear views, as illustrated in the expanded

box at the bottom of Figure 3. The 45-degree gait models differ from the side models in having less occlusion of the arms and legs in the minimum extent configuration and a smaller spread of the arms and legs in the maximum extent configuration. For the front and back views the gait model consists of just a single state because the positions of the arms and legs do not change appreciably during the walking cycle from these viewpoints.

The full walking model thus has 26 states, four states for each of the side views (left and right), four states for each of the 45-degree views, and one state each for the front and back views. Conceptually there are three different types of transition edges in this graph: *self transitions* from a given representative view to itself, *gait transitions* from one state in a gait cycle to the next but maintaining the same viewpoint, and *view transitions* where the viewpoint changes but the state in the gait cycle stays the same. One could further imagine a combined transition that simultaneously goes from one viewpoint to the next and one gait state to the next, but we do not consider that here.

3.2. Estimating Model Parameters

Given the graph structure there are a number of parameters to be estimated. We use the maximum likelihood estimation approach in [4] to learn pictorial structure spatial models from a set of labeled training images. Each training image specifies the viewpoint and the locations of the ten model parts (torso, head, and upper and lower left and right arms and legs). All spatial relations between the parts are learned from the examples, including covariances capturing the degree of change in the relative locations of the parts for a given representative view. In practice many of the viewpoints have the similar spatial relations between their parts, so training need not be done separately for each viewpoint. For instance in the 26 view model above there are only 5 distinct spatial arrangements of parts: the minimum and maximum extent side views, the minimum and maximum extent 45 degree views, and the front/back view.

For the transition probabilities in the HMM we simply use constants corresponding to each of the three transition types identified above: τ_s for self transitions, τ_g for gait transitions and τ_v for view transitions. The transition probabilities for a given node are then obtained by normalizing these values so that the total over all outbound edges from the node is 1.0. We set these constants such that $\tau_s > \tau_g > \tau_v$, and use the same values for all experiments.

In order to generate more accurate locations of occluded parts we interpolate the pictorial structure model parameters between the representative view states, as discussed above. We use linear interpolation of the parameters, although for walking a more accurate model would reflect that near the representative views the change is slower than halfway between such views. The number of intermediate views is determined by the rate of walking. In principle this can be estimated from sample video sequences, but we simply use a fixed value at the moment.

This modeling approach provides a framework for learning parameters such as the transition probabilities. Thus possible extensions to this work include clustering images to automatically identify distinct views, and estimation of HMM parameters using EM.

3.3. Approximations

While pictorial structure matching is efficient given the large number of parameters to be estimated, it still takes approximately half a minute to match a ten-part person model to an image using a 2 GHz processor. With one pictorial structure model per state, the processing requirements for a large spatio-temporal model such as the one here is thus many minutes per frame (although quite easily parallelized). In this section we consider two methods for speeding up processing in the specific case of tracking walking people. The resulting technique runs at about 15 seconds per frame for the full model considered above (on a 2.0 GHz Pentium 4). The first technique uses the horizontal variance of the silhouette pixel locations to determine the best state sequence, and then only matches the pictorial structure model that corresponds to the best state at each time. The second technique uses a bounded velocity assumption to limit the parameter range that is searched in finding the best pictorial structure parameters for a given image frame.

Our goal here is to compute a best state sequence $(Q_1^*, \ldots, Q_T^*)$ without needing to estimate the integral in (2) for every model and image. Thus we consider an approximation to the observation probabilities $P(Y_t = k|Q_t = i)$. This approximation is based on using the horizontal variance of the model pixel locations rather than the full model. As has been observed in other work (e.g., [2]), most of the change in the silhouette of a walking person is in the horizontal locations of the pixels. Let $f(t)$ denote the variance in the x-locations of the silhouette pixels at time t, where the silhouette has first been scaled to a standard height in order to accommodate overall size changes. We define the normalized horizontal variance as the difference between $f(t)$ and a smoothed version of $f(t)$. Namely, $g(t) = f(t) - (f(t) \star G_p(t))$, where G_p is a Gaussian of standard deviation p, the period of $f(t)$. The period, p, can be determined easily using auto-correlation or the FFT. If the speed of walking changes substantially then windowing can be used to vary the smoothing.

An example of the normalized horizontal variance is shown in Figure 4 for a video sequence of a person walking

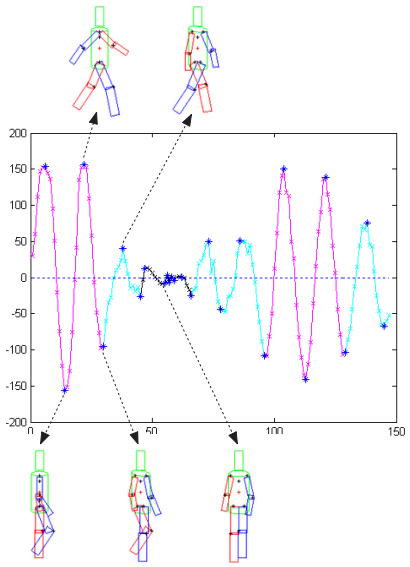

Figure 4: *The normalized horizontal variance, $g(t)$, for a person walking in a circle, illustrating the combined effects of changes in viewpoint and gait cycle.*

around in a circle, so the viewpoint is changing. The Figure also shows corresponding representative pictorial structure models for several of the views. Note that the primary effect of the normalization is to create a common "zero point" around which $g(t)$ oscillates in a periodic manner. The gait cycle is captured by the oscillation of the function and the viewing direction is captured by its amplitude. In order to explicitly represent the amplitude at each time we also compute the separation between the minimal and maximal values over each period, $h(t) = g_{\max}(t) - g_{\min}(t)$ where the min and max are computed for the period containing t.

For states within a gait cycle, the horizontal variance $g(t)$ distinguishes the two minimal extent states, which have negative values of $g(t)$, from the two maximal extent states, which have positive values. Analogously to the pictorial structure models, in a single time frame the two minimal extent states cannot be distinguished from one another, nor can the two maximal extent states, but these states are distinguishable based on transitions between states over time.

For changes in viewpoint, the amplitude $h(t)$ distinguishes the two side views (colored magenta in the Figure), which have high amplitude, from the four 45 degree views (cyan), which have moderate amplitude, from the front and back views (black), which have low amplitude. However, even with the associated state transitions, $h(t)$ is not enough to distinguish different 45 degree views. Recall the view transition diagram in Figure 3. The transitions from a side view (or from a front/back view) are both to 45 degree views which have essentially the same amplitude. However, while the overall amplitude is the same, these two views differ in terms of whether or not the 45 degree viewing direction aligns both the forward and backward arms with the torso (and vice versa for the legs as they are in opposing positions to the arms) while the viewpoint change is happening. Thus we also compute the amplitude of the normalized horizontal variance separately for the upper half of the silhouette (corresponding to the arms) and the lower half (corresponding to the legs) to capture these differences. Call these amplitudes h^u and h^l, respectively.

We use a normal distribution to model these parameters for a given state i. Thus there are a total of eight parameters to this model, the mean and variance for each of the normalized horizontal variance g, its amplitude, h, and the amplitudes for the upper and lower halves h^u and h^l. Let v_i denote these eight parameters for state i. We approximate the observation probabilities using $P(Y_t = k|v_i)$. Given the best state sequence $Q^* = (Q_1^*, \ldots, Q_T^*)$ found using this approximation, we then compute the spatial parameters Θ_t for each time frame by matching the pictorial structure model corresponding to state Q_t^* to the image at time t by sampling from the distribution in (1). As discussed above, the pictorial structure model is interpolated between representative views, so as to provide a stronger prior on the locations of parts and thus increase the localization accuracy for occluded parts.

In the final step of computing the spatial parameters Θ_t, we use a bounded velocity assumption to further speed up matching by eliminating large portions of the parameter space from consideration when sampling the posterior distribution of model parameters in equation (1). The location of each part is characterized by a set of parameters. Bounding the velocity implies bounds on the degree of change in the parameters from one time frame to the next. This makes it possible to restrict the domain over which the posterior must be estimated. We use a simple hypercube, limiting each parameter to an interval around the corresponding value for the previous time-frame. Recall that estimating the posterior involves computing both the likelihood and prior over possible values of these parameters for each part, whose domains are in turn limited. This results in an order of magnitude speedup. Such windowing cannot only drastically shrink the volume of the state space that needs to be computed, it can also improve the accuracy of the matching results when there is a good prior on the part locations but certain parts are hidden from view.

3.4. Examples

In this section we present some experimental results to demonstrate the capabilities of the method in tracking a human silhouette. We consider three sequences, the first two have a fixed viewpoint and in the third a person walks in a circular path so that the viewpoint changes extensively.

The first two sequences are from the multi-view imagery

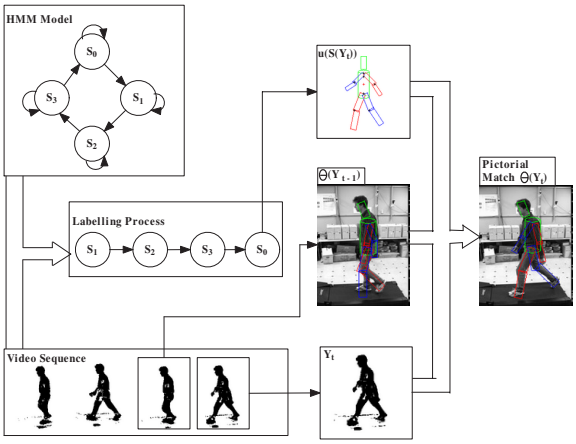

Figure 5: *This graph illustrates our experimental framework. The three major steps are, (1) silhouette extraction, (2) state labeling using approximation, (3) windowed-sampling based pictorial structure matching.*

in the CMU HID dataset. Each sequence is an 11 second clip of a person walking on a treadmill. One sequence is taken from a side-view and the other from a front-side 45 degree view. The third sequence, from Michael Black, is of a person walking in a circle and poses substantial challenges in terms of self occlusion and viewpoint change, demonstrating the power of having integrated spatial and temporal models (we refer to this as the PARC sequence).

Our experimental framework can be summarized by the illustration in Figure 5. There are three major steps in the whole process. The first step is silhouette extraction using background subtraction because of the stationary camera. The remaining steps of processing are all covered in the previous sections. The second major processing step involves calculating the horizontal variance measures and Viterbi estimation of the best state sequence given the walking model and the horizontal variance. The third major step is estimation of the model parameters for the pictorial structure model corresponding to the best state at each time t (including the use of interpolated models between representative views). These model parameters are then used to overlay the model on the original images. The calculations are all done using a size-normalized coordinate frame and then mapped back into image coordinates.

Figure 6 shows some of the results for the side-view walking sequence. These frames illustrate that our approach can handle self-occlusion and the ambiguity in the identity of the parts in individual frames. Note that the distinctions between the left (colored red) and right (colored blue) limbs are correctly tracked over time, including through occlusions. The matching is completely automatic, except that for the first frame we specify whether the left arm and right leg are forward, or vice versa, as this cannot be distinguished from the image data. This is done by setting the initial state

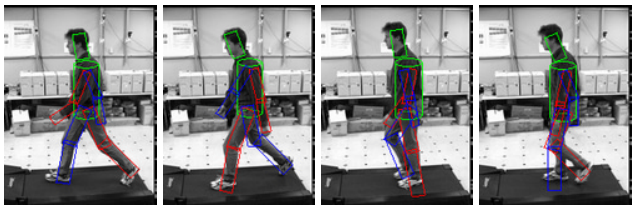

Figure 6: *Results for the side view of the treadmill sequence.*

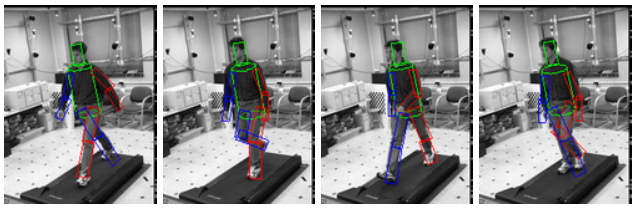

Figure 7: *Results for the 45-degree view of the treadmill sequence.*

probabilities of the HMM. Figure 7 shows some results for the 45-degree viewpoint sequence, again correctly tracking the parts over time and through occlusions. These two sequences illustrate the capabilities of the single-view four-state HMM gait model, using pictorial structures for the states.

Figure 8 shows the results of our method for the PARC sequence, where a person walks in a circle. Note the substantial changes in viewpoint and self occlusions. This result demonstrates that the full model, incorporating both viewpoint and gait changes, can successfully track the left and right side limbs over a sequence that combines many successive changes in viewpoint simultaneous with the self-occlusions due to the gait cycle. This is a challenging sequence and we are not aware of other methods that can successfully track the parts over time as demonstrated here.

4. Conclusions

We have presented a modeling framework that integrates pictorial structure spatial models with hidden Markov temporal models, and have shown an efficient inference procedure for finding a maximum a posteriori (MAP) probability state sequence for such a model. Our method also produces a sequence of configuration parameters corresponding to a high posterior probability match of the model at each image frame. We then described an approximation that uses the horizontal variation of pixel locations to more efficiently determine a MAP state sequence. The results illustrate the ability of this combined spatio-temporal modeling technique to correctly track the position and identity of a person's limbs through self occlusions and changes in viewing direction.

References

[1] C. Bregler, Learning and Recognizing Human Dynamics in Video Sequences. *CVPR*, San Juan, Puerto Rico, June 1997

[2] R.T. Collins, R. Gross and J. Shi, Silhouette-Based Human Identification from Body Shape and Gait, International Conference on Face and Gesture, pp. 351-356, 2002.

[3] A. Elgammal, V. Shet, Y. Yacoob and L.S. Davis, Learning Dynamics for Exemplar-Based Gesture Recognition. *CVPR*, Madison, Wisconsin, June 2003.

[4] P.F. Felzenszwalb, D.P. Huttenlocher, Pictorial Structures for Object Recognition, IJCV to appear.

[5] M.A. Fischler and R.A. Elschlager. The representation and matching of pictorial structures. TC, 22(1):67-92, January 1973.

[6] D. M. Gavrila. The Visual Analysis of Human Movement: A Survey. Computer Vision and Image Understanding, Academic Press, vol. 73, nr. 1, pp. 82-98, 1999.

[7] S. X. Ju, M. J. Black, and Y. Yacoob. Cardboard people: A parameterized model of articulated motion. In 2nd Int. Conf. on Automatic Face and Gesture Recognition, pages 38-44, 1996.

[8] V. Pavlovic, J. M. Rehg, and J. MacCormick. Learning switching linear models of human motion. *NIPS*, 2000.

[9] L. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE* Vol. 77(2), pp. 257-286, 1989.

[10] D. Ramanan and D. A. Forsyth. Finding and tracking people from the bottom up. *CVPR*, Madison, Wisconsin, June 2003.

[11] B. Stenger, A. Thayananthan, P.H.S. Torr, and R. Cipolla. Filtering Using a Tree-Based Estimator. *ICCV*, Vol. II, pages 1063-1070, Nice, France, October 2003

[12] A.D. Wilson and A.F. Bobick, Parametric Hidden Markov Models for Gesture Recognition, *IEEE Trans. PAMI*, v. 21, pp. 884-900, 1999.

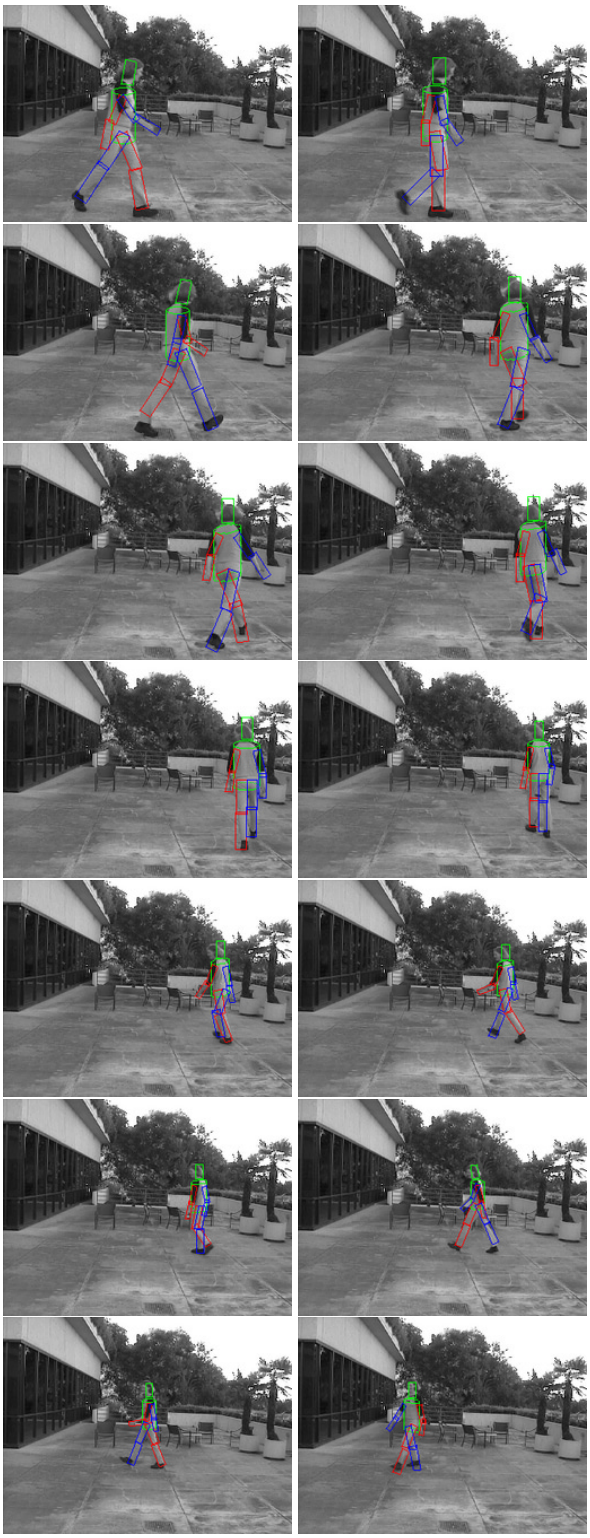

Figure 8: *Results for the multi-viewpoint sequence.*