Combining Spatial and Temporal Priors for Articulated Human Tracking with Online Learning

Cheng Chen and Guoliang Fan
School of Electrical and Computer Engineering
Oklahoma State University
Stillwater, Oklahoma, USA
guoliang.fan@okstate.edu

Abstract

We study articulated human tracking by combining spatial and temporal priors in an integrated online learning and inference framework, where body parts can be localized and segmented simultaneously. The temporal prior is represented by the motion trajectory in a low dimensional latent space learned from tracking history, and it predicts the configuration of each body part for the next frame. The spatial prior is encoded by a star-structured graphical model and embedded in the temporal prior, and it can be constructed "on-the-fly" from the predicted pose and used to evaluate and correct the prediction by assembling part detection results. Both temporal and spatial priors can be online learned incrementally through the Back Constrained-Gaussian Process Latent Variable Model (BC-GPLVM) that involves a temporal sliding window for online learning. Experiments show that the proposed algorithm can achieve accurate and robust tracking results for different walking subjects with significant appearance and motion variability.

1. Introduction

Markerless human tracking and pose estimation have recently attracted intensive attention in the computer vision community because of their wide applications. Meanwhile they also remain to be one of the most challenging research issues largely due to the ambiguity, complexity and nonlinearities in observed video sequences. Using appropriate prior knowledge (such as motion or shape) would make the problem better defined and hopefully easier to tackle. The major advantages of using priors are to reduce the search space by taking advantages of various constraints and to ensure a plausible solution. Two commonly used priors are spatial and temporal priors, both of which play important roles for human detection and tracking, and have been well studied by many researchers in different context.

Spatial priors are usually defined on body parts and characterize the spatial configuration of a certain pose [3]. One important question is how to make one spatial prior adaptable to a large number of pose variations. One straightforward extension is to train separate spatial priors for several typical poses [9]. However, a spatial prior representation that can only handle a discrete pose variable has difficulty to characterize the smooth and continuous pose transition in a video sequence. On the other hand, temporal priors specify certain dynamic constraint of human motion [19], and they can ensure the temporal continuation across adjacent poses. Most temporal models do not impose a strong spatial constraint among body parts or treat each part independently for tracking [18]. How to learn the two priors are also of great interest. There are two kinds of learning strategies. *Off-line learning* normally requires sufficient and/or diverse training samples and usually leads to the learned priors that favor the training data. *Online learning* can learn the priors "on-the-fly" that is more favorable and effective to deal with human motion with significant variability even from different activities [17].

In this work, we propose a new framework for articulated human tracking that integrates both spatial and temporal priors and is supported by online learning. The idea of combining both priors has been well acknowledged and incorporated into most tracking algorithms where both priors are usually learned off-line and one prior often overshadows the other one during inference. In our work, the spatial prior is embedded in the temporal prior, and both priors are learned online from past tracking history in an incremental way. Specifically, the temporal prior can predict the pose for the next frame that induces a pose specific spatial prior. This spatial prior in return is used to evaluate and correct the pose prediction by assembling part-level detection. Our approach distinguishes itself from others in that it incorporates both online learned spatial and temporal priors in one integrated inference framework.

2. Related Works

The biological vision model proposed in [16, 6] suggested two perception pathways in motion perception, the *appearance pathway* and the *motion pathway*. It was considered as one of the major breakthroughs in recent vision research [1], and motivates researchers to involve both spatial (appearance) and temporal (motion) priors in their tracking algorithms. Broadly speaking, related work can be classified into two groups: the *temporal-prior dominated approaches* and the *spatial-prior dominated approaches*.

In [9], a unified spatial-temporal articulated model was proposed for human tracking, where the pose is a discrete variable and defined as the hidden state of a hidden Markov model (HMM). The temporal prior is incorporated as a state transition matrix, and then the tracking task is formulated as a Bayesian estimation problem. In [15], a single pictorial structure graph model was extended into a dynamic Bayesian network (DBN), where the probabilistic relationships between joints at a given time instant as well as those over time can be learned from motion capture data. Then belief propagation is used as the inference engine to effectively incorporate the top-down spatial prior with bottom-up part detection for articulated human tracking. Along the same venue, a temporal pictorial structure model was developed in [13], which mainly relies on appearance priors for human tracking. Above methods are considered as the *spatial-prior dominated* ones where the spatial prior plays a more important role and only weak temporal priors are involved for dealing with activity variation.

The human pose can be represented in a high dimensional (HD) parameter space where the distribution of plausible human poses is very sparse. Various non-linear dimensionality reduction (DR) techniques were proposed to explore the low-dimensional (LD) intrinsic structures for a compact pose representation. The Gaussian Process Latent Variable Model (GPLVM) [10] is an effective DR technique that offers a smooth mapping from the LD latent space to the HD kinematic space. Several GPLVM variants were developed for temporal series analysis. For examples, Gaussian Processing Dynamic Models (GPDM) [19] were specifically designed for human motion tracking by introducing a dynamic model on the latent variable that can be used to produce tracking hypothesis in a latent space[18]. Back Constrained-GPLVM (BC-GPLVM) [11] improves the continuity in the latent space by enforcing the local proximities in both the LD and HD spaces. Consequentially, BC-GPLVM produces a smooth motion trajectory in the latent space that can be used as a non-parametric dynamic model for human tracking [8]. All of above DR methods focus on the exploration and exploitation of temporal priors of human motion, and they do not involve spatial (kinematic) priors explicitly. Therefore, we consider them as temporal-prior dominated approaches.

Motivated by previous research, we want to take advantage of the complementary nature of the above two methodologies. On the one hand, our work is similar to [9, 13] in the sense of how the spatial prior is represented. But we involve a *strong* temporal prior that can handle a continuous pose variable. On the other hand, our algorithm inherits some ideas from [8, 14] regarding how the temporal prior is developed for top-down prediction. However, we use a *structured* spatial prior that fuses part detection results to evaluate and correct the prediction. Moreover, we explore the synergy between the two priors in the context of online learning, which is inspired by the local mixed Gaussian process regressors proposed in [17]. To the best of our knowledge, there is no prior research on how to combine spatial and temporal priors in an online learning framework.

3. Background Review

We firstly briefly review the two major building blocks regarding the representations of temporal and spatial priors.

3.1. Temporal Prior Modeling: GPLVM

The Gaussian process latent variable model (GPLVM) [10] is an effective method to learn $X = \{\mathbf{x}_i\}_{i=1}^N, \mathbf{x}_i \in R^q$ in a LD latent space from $Y = \{\mathbf{y}_i\}_{i=1}^N, \mathbf{y}_i \in R^D$ $(D >> q)$ in a HD observation space, and it also provides a probabilistic mapping from X to Y. We refer the readers to [10, 5] for more details. Assuming each observed data point, $\mathbf{y}_i$ is generated through a noisy process from a latent variable $\mathbf{x}_i$,

$$\mathbf{y}_i = f(\mathbf{x}_i) + \epsilon, \tag{1}$$

where $\epsilon \sim N(0, \beta^{-1}I)$. Assuming a Gaussian distribution over functions $f \sim N(0, k(\mathbf{x}_i, \mathbf{x}_j))$. The covariance $k(\mathbf{x}_i, \mathbf{x}_j)$ characterizes the nature of the functions. One widely used covariance function is

$$k(\mathbf{x}_i, \mathbf{x}_j) = \theta_1 e^{-\frac{\theta_2}{2}\|\mathbf{x}_i - \mathbf{x}_j\|^2} + \theta_3 + \beta^{-1}\delta_{i,j}, \tag{2}$$

where the parameters are given by $\Phi = \{\theta_1, \theta_2, \theta_3, \beta\}$ and $\delta_{i,j}$ is the Kronecker's delta function. The scalar $k(\mathbf{x}_i, \mathbf{x}_j)$ models the proximity between two points $\mathbf{x}_i$ and $\mathbf{x}_j$.

After GPLVM learning, given a new latent variable $\mathbf{x}_*$, the likelihood of the corresponding HD data point $\mathbf{y}_*$ is:

$$p(\mathbf{y}_*|X, \mathbf{x}_*) = N(\mathbf{y}_*|\mu, \sigma^2), \tag{3}$$

where

$$\mu = Y^T K_{X,X}^{-1} \mathbf{k}_{X,\mathbf{x}_*}, \tag{4}$$

where $K_{X,X} = \{K_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)\}$, and $\mathbf{k}_{X,\mathbf{x}_*}$ is a column vector developed from computing the elements of the kernel matrix between the learn latent state data X and the new point $\mathbf{x}_*$. The variance that is then given below will increase as $\mathbf{x}_*$ deviates from the training data X.

$$\sigma^2 = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_{X,\mathbf{x}_*}^T K_{X,X}^{-1} \mathbf{k}_{X,\mathbf{x}_*}. \tag{5}$$

To ensure a smooth trajectory in the latent state space for temporal series data, BC-GPLVM was proposed in [11] that enforces local proximities in both the LD and HD spaces. In our work, BC-GPLVM is used to learn a compact LD representation of human motion in the latent space and a probabilistic reverse mapping from the LD latent space to the HD observation space. We adopt the BC-GPLVM to a local online learning strategy [17].

3.2. Spatial prior: Star-structured Graphic Model

The pictorial structure based spatial prior representation has become an increasingly compelling approach for articulated human body tracking. Following [3], we represent the spatial prior for a pose by a star-structured graphical model Ψ. Let us regard pose $\mathbf{y} = (l_1, ..., l_d)$ as a vector of 2D configuration (position and orientation) of d body parts. The joint distribution of d part configuration with respect to pose $\mathbf{y}$ can be written as the following:

$$p_\Psi(\mathbf{y}) = p_\Psi(l_1, ..., l_d) = p_\Psi(l_r) \prod_{k \neq r} p_\Psi(l_k|l_r), \quad (6)$$

where l_k and l_r are the configuration parameters for non-reference part k and the reference part r respectively. By assuming the conditional probability density functions for $p_\Psi(l_k|l_r)$ following the Gaussian distribution. Then, for each non-reference part k, the conditional distribution of its configuration with respect to pose Ψ is defined below,

$$p_\Psi(l_k|l_r) = \mathcal{N}(l_k - l_r|\mu_k, \Sigma_k). \quad (7)$$

We can also assume a Gaussian distribution for $p(l_r)$.

$$p_\Psi(l_r) = \mathcal{N}(l_r|\mu_r, \Sigma_r). \quad (8)$$

In an off-line learning framework, the parameters of the star model are often obtained by a maximum-likelihood estimator (MLE) from the labeled training data. For a test image, this spatial prior is used to assembly part detection results, or called *map* images, which indicate the confidence of the existence of each part at every pixel location. Edge histogram-based part detection was used in [3] where a distance transform-based fast inference algorithm is also developed to assemble *map* images for detection and localization. In [2], a segmentation-based hypothesis-and-test method was proposed to produce more salient *map* images for part detection that improves the whole-part localization accuracy. We will make two extensions to the star model-based spatial prior in this work. (1) The spatial prior is *time variant* and is able to handle a continuous pose variable rather than a discrete pose variable in [9, 2]. (2) The spatial prior is embedded in the temporal prior, and can be constructed "on-the-fly" based on the temporal prediction for every incoming frame rather than learned offline [3].

4. Proposed Research

4.1. Research Overview

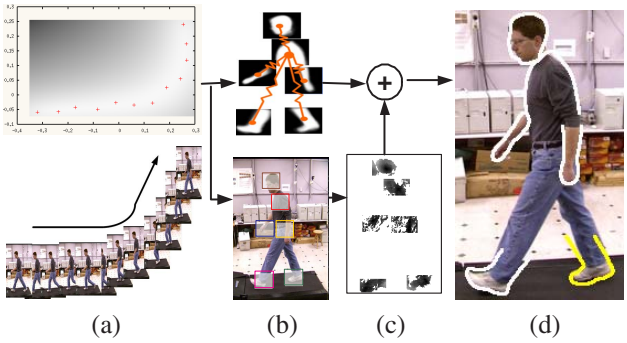

Figure 1. The algorithm flow. (a) Online learning and dynamic prediction in the latent space. (b) Pose prediction in the observation space and construction of the star model. (c) Local part detection according to the prediction. (d) Localization by assembling the part detection results via the star model.

Our algorithm is featured by the marriage of two powerful mathematical tools, BC-GPLVM and the star-structured graphical model, which is elaborated in the context of online learning. The synergy between the two priors is explored by embedding the spatial prior into the temporal prior and learning them together. The proposed algorithm involves four major steps as follows.

- *Online learning and Pose Prediction:* From past tracking history, we learn a smooth motion trajectory in the latent space via BC-GPLVM, as shown in Fig. 1(a), which can be used as a non-parametric dynamic model to predict the next pose in the latent space by B-spine extrapolation.

- *Spatial Prior Construction:* Based on the prediction in the latent space, we can predict the next pose in the HD observation space via the LD-HD reverse mapping, as shown in Fig. 1(b). The predicted pose specifies the possible location of each body part in the next frame that enables the efficient local search of body parts. Also, a star model is constructed accordingly to represent the pose specific spatial prior.

- *Local Part Detection:* Based on the pose prediction, local part detection is performed for d (the number of body parts) body parts that results in d localized *map* images that are shown together in Fig. 1(c).

- *Pose Correction:* The pose specific star model is used to assemble the part detection outputs and produce the final localization results for the whole body as well as body parts, as shown Fig. 1(d).

4.2. Problem Formulation

Given N image frames $I_{1:N}$, we want to estimate the pose $\mathbf{y}_i = (l_1^{(i)}, ..., l_d^{(i)})$ for each frame where $\mathbf{y}_i$ is a vector of 2D configuration (position and orientation) of d body parts at frame i. Let $\mathbf{x}_i$ to be the latent state associated with $\mathbf{y}_i$. Let $P_i = \{p_1^{(i)}, ..., p_d^{(i)}\}$ be the appearance models (e.g., an object template) of the d body parts. Given the pose estimation results of N previous frames, i.e., $Y_{1:N} = \{\mathbf{y}_1, ..., \mathbf{y}_N\}$, current body part models p_N and next frame I_{N+1}, part localization (tracking) results can be obtained by maximize the posterior probability:

$$\mathbf{y}^*_{N+1} = \arg\max_{\mathbf{y}_{N+1}} P(\mathbf{y}_{N+1}|I_{N+1}, P_N, \mathbf{y}_{1:N}). \quad (9)$$

Generally, it is intractable to find the $\mathbf{y}^*_{N+1}$ directly due to its HD nature. Hence we use a *prediction-and-correction* framework to attack this problem, as shown in Fig. 2.

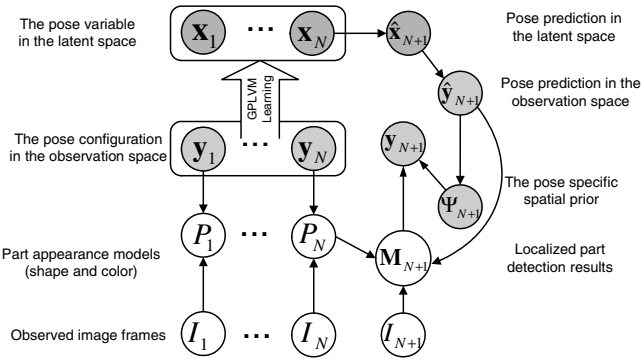

Figure 2. Problem formulation by a graphical model.

Assume we can learn a smooth motion trajectory $X_{1:N} = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$ in the latent space via BC-GPLVM based on past tracking history $Y_{1:N}$. We can predict the next pose in the latent space first, $\hat{\mathbf{x}}_{N+1}$, which can be converted to $\hat{\mathbf{y}}_{N+1}$ in the image space. $\hat{\mathbf{y}}_{N+1}$ has two implications. First, it can be used to produce d localized *map* images, $\mathbf{M}^{(N+1)}_{maps} = \{M_1^{(N+1)}, ..., M_d^{(N+1)}\}$. Second, it defines a pose specific spatial prior represented by a star model Ψ_{N+1}. Following the same idea as in [3], these *map* images can be assembled by Ψ_{N+1} in the form of star-structured graphical model. Then the tracking problem can be reformulated as maximizing the posterior probability:

$$\mathbf{y}^*_{N+1} = \arg\max_{\mathbf{y}_{N+1}} P(\mathbf{y}_{N+1}|\mathbf{M}^{(N+1)}_{maps}, \Psi_{N+1}), \quad (10)$$

where Ψ_{N+1} is the spatial prior represented in (6). The optimization problem of (10) can be efficiently solved using the fast inference algorithm developed in [3]. Then $\mathbf{y}_{N+1}$ can be used to achieve the updated appearance models P_{N+1} based on I_{N+1}, and will be involved in the next step BC-GPLVM learning to predict $\mathbf{y}_{N+2}$ as the temporal sliding window moves forward one frame.

5. Learning and Inference

In this section, we detail the four major steps for learning and inference in our tracking algorithm.

5.1. Online Learning and Dynamic Prediction

In general, a pose can be represented by a HD vector that records joint angles or positions. Here we use a simple body representation with six body parts where each part is specified by the 2D position and orientation in the image domain. Given a pose series $Y_{1:N}$, BC-GPLVM can be used to learn the kernel parameters $\Phi = \{\theta_1, \theta_2, \theta_3, \beta\}$ and the latent variable series $X_{1:N}$. Different from off-line learning, we use a temporal sliding window to involve recently estimated poses for online (local) BC-GPLVM learning. The learned model is only used once for pose prediction in the next frame. As shown in Fig. 3, although there is no explicit dynamic model involved in BC-GPLVM, the temporal constraint is well-reflected by the smooth motion trajectory in the LD latent space. We can extrapolate this motion trajectory to predict the pose in the next frame.

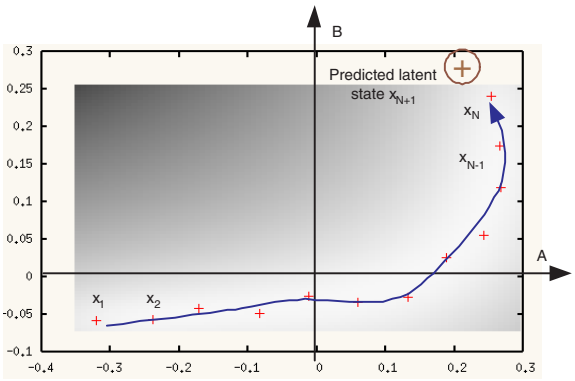

Figure 3. An example of BC-GPLVM online learning and dynamic prediction via B-spline extrapolation in the 2D latent space.

Let $\mathbf{x}_i = (a_i, b_i)^T$, we can apply the B-spline regression process on the latent states $X = \{\mathbf{x}_i\}_{i=1}^N$, as shown in Fig. 3. The two obtained B-spline functions $\mathcal{A}(\cdot)$ and $\mathcal{B}(\cdot)$ will satisfy the following constraints:

$$\begin{cases} \mathcal{A}(i) \cong a_i, \\ \mathcal{B}(i) \cong b_i, . \end{cases} \quad (11)$$

where $i = 1, ..., N$. Then through B-spline extrapolation $a_{N+1} = \mathcal{A}(N+1)$ and $b_{N+1} = \mathcal{B}(N+1)$, we can compute the predicted latent state for the next frame $(N+1)$ as $\mathbf{x}_{N+1} = (a_{N+1}, b_{N+1})^T$, (as indicated by the circled marker in Fig. 3). From the predicted latent variable $\mathbf{x}_{N+1}$, the associated pose in the image space $\mathbf{y}_{N+1}$ can be constructed through the reverse LD-HD mapping given in (4), and defined as:

$$\hat{\mathbf{y}}_{N+1} = Y^T K_{X,X}^{-1} k_{X,\mathbf{x}_{N+1}}. \quad (12)$$

5.2. Constructing the Spatial Prior

The uncertainty of $\hat{\mathbf{y}}_{N+1}$ is reflected by the variance defined in (5). So far, it is assumed that the configurations of d parts are independent, indicating a weak spatial constraint if only the temporal prior is used. In order to incorporate the spatial constraint among body parts, we need to construct the pose specific spatial prior represented by the star model from $\hat{\mathbf{y}}_{N+1}$. That means we need to estimate the conditional distributions between each non-reference part and the reference part, which can be derived straightforwardly from the Gaussian assumption of $\hat{\mathbf{y}}_{N+1}$. Therefore, the conditional distribution defined in (7) will become:

$$p_\Psi(l_k|l_r) = \mathcal{N}(l_k - l_r|\mathbf{y}^k_{(N+1)} - \mathbf{y}^r_{(N+1)}, 2\sigma^2 \cdot \mathbf{I}), \quad (13)$$

where σ^2 is given (5), $\mathbf{y}^r_{(N+1)}$ is the configuration of the reference part, and $\mathbf{y}^k_{(N+1)}$ is the relative configuration of non-reference part k with respect to the reference part. Similar to (8), the distribution of the reference part will become:

$$p_\Psi(l_r) = \mathcal{N}(l_r|\mathbf{y}^r_{(N+1)}, \sigma^2 \cdot \mathbf{I})). \quad (14)$$

Strictly speaking, the covariance matrices of (13) and (14) need to be added with an additional error term to accommodate the prediction error in the latent space. In this work, the value of this prediction error term is set by experiment, and it is found that the tracking performance is not very sensitive to the choice of this value.

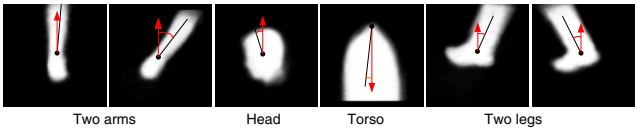

Figure 4. Off-line leaned part shape models where the average orientation of each part is also given.

5.3. Part Detection

The predicted pose $\hat{\mathbf{y}}_{N+1}$ specifies the possible locations of d body parts that could define a search region for each part. Ideally, each search region should be isotropic and determined by (5). For simplicity, we use a square region of 21×21 for local part detection which is centered with the part position encoded in $\hat{\mathbf{y}}_{N+1}$. Similar to [2], we resort to a segmentation-based hypothesis-and-test method for part detection where off-line learned shape models are used, as shown in Fig. 4. These shape models can be further represented by the coupled region-edge shape priors which are used to compute *map* images given an image represented by watershed cells. At each hypothesized location, the region prior is used to form a segmentation by merging watershed cells, while the edge prior is applied to evaluate the formed segmentation in terms of *shape similarity* and *boundary smoothness*.

To take advantage of localization results in the previous frame, we modify the evaluation criterion for computing the *map* images by replacing the *boundary smoothness* with the *template matching score*. Given a segmentation Z formed in a location, we represent the boundary of Z by $\Gamma(Z)$, and the new evaluation score for Z is given by

$$\rho_{\mathcal{M}}(Z) = -d(\Gamma(Z), \mathcal{M})) + \zeta SAD(I_{(N+1)}, P_N), \quad (15)$$

where the first term is the chamfer distance indicating the shape similarity between $\Gamma(Z)$ and the off-line learned edge prior $\mathcal{M}$, and the second term is the *SAD* (Sum of absolute differences) that reflects the degree of match between the online learned template P_N and $I_{(N+1)}$. ζ balances the relative importance between the off-line and online learned part priors. Some part detection examples are shown in Fig. 5(b) where a dark pixel value indicates a high possibility of the existence of a part. The computation of these *map* images can be very efficient due to local part detection constrained by $\hat{\mathbf{y}}_{N+1}$, instead of the full search used in [3, 2].

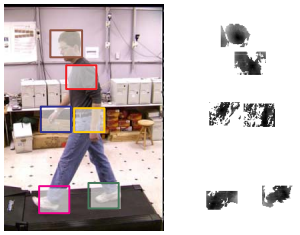

Figure 5. Part detection results. (a) An image with predicted local search regions. (b) The localized *map* images for six body parts.

5.4. Pose Correction

The part-based *map* images will be assembled by the online learned spatial prior through the star-structured graphical model. The same as in [3], given the set of *map* images $\mathbf{M}^{(N+1)}_{map} = \{M^{(N+1)}_k | k = 1, ..., d\}$, the optimal $\mathbf{y}_{(N+1)}$ can be obtained by re-writing (10) as,

$$\mathbf{y}^*_{(N+1)} = \arg\max_{\mathbf{y}_{(N+1)}} p_\Psi(\mathbf{y}_{(N+1)}|\mathbf{M}^{(N+1)}_{maps}). \quad (16)$$

Using Bayes law,

$$p_\Psi(\mathbf{y}_{(N+1)}|\mathbf{M}^{(N+1)}_{maps}) \propto p_\Psi(\mathbf{M}^{(N+1)}_{maps}|\mathbf{y}_{(N+1)}), p_\Psi(\mathbf{y}_{(N+1)}). \quad (17)$$

where

$$P_\Psi(\mathbf{M}^{(N+1)}_{maps}|\mathbf{y}_{(N+1)}) = \prod_{k=1}^{k=d} M^{(N+1)}_k \left(\mathbf{y}^k_{(N+1)}\right),$$

where $M^{(N+1)}_k \left(\mathbf{y}^k_{(N+1)}\right)$ is the value of the *map* image of part k in location $\mathbf{y}^k_{(N+1)}$. Also $p_\Psi(\mathbf{y}_{(N+1)})$ can be evaluated through the learned star-structured graphical model defined in (6). A fast distance transform based inference method can be used to solve this problem efficiently [3].

6. Occlusion Handling

Occlusion handling is an important issue for articulated human tracking where some body parts may be invisible for some poses. In this work, we are interested self-occlusion, and our method is inspired by the *multi-object tracking theory* proposed in [4], where the notion of "*object files*" was developed to store episodic representations for real-world objects. Each object file contains the joint spatio-temporal information (such as appearance and motion) about a particular object in a scene. An "object file" is established for each body part being tracked that plays an important role for occlusion handling. The algorithm flow is shown in Fig. 6 where three occlusion-related issues are addressed.

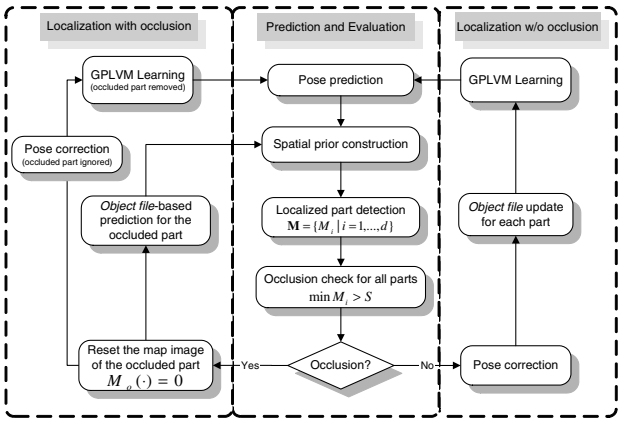

Figure 6. The flow of tracking and occlusion handling.

Occlusion detection: The *map* images can be directly used for occlusion detection. Given a detection threshold S, if $\min M_o(\cdot) > S$, we can declare that part o is occluded. The position estimation of part o only depends on the prior knowledge encoded in the graphical model Ψ.

Prediction of an occluded part: Given part o that is declared in I_N, its "object file" can be used for predicting its position in I_{N+1} as follows:

$$\hat{\mathbf{y}}^o_{(N+1)} = \hat{\mathbf{y}}^r_{(N+1)} + \Delta l_o, \tag{18}$$

where $\hat{\mathbf{y}}^r_{(N+1)}$ is the predicted position of the reference part (assuming the reference part is never occluded), and Δl_o is the relative configuration between part o and the reference part r that can be retrieved from the "object file" of part o.

Learning and inference for occlusion: When part o is occluded in I_N, we disable its *map* image by setting $M_o(.) = 0$. Its configuration (given by the spatial prior) will be ignored in the online BC-GPLVM learning for I_{N+1}.

The above occlusion handling technique is simple yet effective, and could be extended to handle more sophisticated cases. The synergistic use of spatial and temporal priors allows the tracking algorithm to have more flexibility and capability of handling occlusion.

7. Experiment Results

The algorithm was implemented in C++, and the testbed is a PC computer with a Core 2 Duo E4700/2.6GHz CPU and 2GB RAM. We used the CMU Mobo database for algorithm validation [7], which includes video sequences captured from 25 individuals walking on a treadmill. Our algorithm was tested on four side-view sequences: vr03_7, vr08_7, vr21_7 and vr14_7, and we selected a walking cycle of 33 frames for each sequence. The first three have a normal walking style with different appearances (body shapes and colors), and the last one has an abnormal pattern where the subject touched his nose during walking. We have three specific experiments. The first one evaluates the tracking accuracy that is measured by the localization accuracy of each body part over a complete cycle. The second shows the capability of handling an unseen activity with an occluded body part. The last one presents body part segmentation that is part of online appearance learning.

The size of the temporal sliding window for training sample selection has to be determined in practice. We found that a number between 5-12 frames is acceptable. It takes about 200 ms for BC-GPLVM training (100 iterations) over 12 frames. Part-based evaluation is about 200 ms per frame that include the localization and segmentation of each body part. After the initialization on the first 5-12 frames, the proposed tracking algorithm can run at about 2 fps for the following frames.

7.1. Part localization

To evaluate the accuracy of articulated human tracking, we have manually obtained the ground-truth (position and orientation) of six body parts in all test video frames, i.e., the *Head*, *Torso*, *Right_arm* (R_arm), *Left_arm* (L_arm), *Right_leg* (R_leg), *Left_leg* (L_leg). Similar to [15], we evaluate the tracking accuracy by comparing the estimated position/orientation with the ground-truth ones. There are two competing algorithms both of which involve the same part-bard spatial prior that were trained off-line for each typical pose (i.e., *High-point*, *Contact* and *Passing*). One is the *1-fan* method [3] that involves the edge histogram for part detection, and the other is the *hybrid* approach [2] where coupled edge-region shape priors are used for part detection. Both algorithms require several pose specific spatial priors that are learned off line, and they do not involve any temporal prior by treating each frame independently. Although the hybrid approach improves the localization accuracy for six body parts compared with the 1-fan method, it is time-consuming due to the fact that segmentation is involved in part detection. The proposed algorithm is much more efficient and effective because the combined spatio-temporal priors are online learnt for each pose (a continuous variable) and dramatically narrows the local search for part detection.

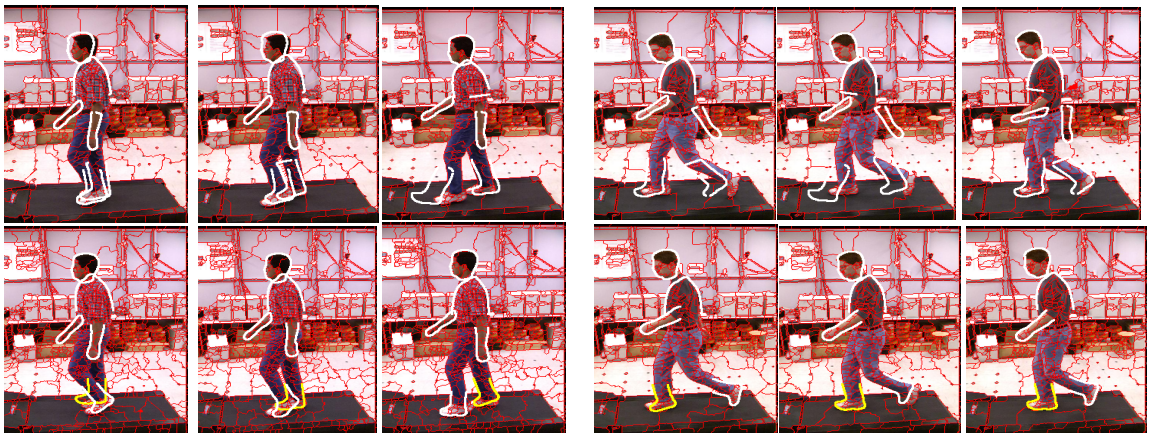

Figure 7. The comparison between the *hybrid* approach [2] (top) and the proposed one (bottom) for three frames from two test videos.

The comparative results are shown in Table 1, where the results from the 1-fan and hybrid approaches are the averages over three typical poses, while that of ours is the average over a complete walking cycle. Our algorithm demonstrates significant improvements over the two competing algorithms in tracking accuracy and efficiency. The localization performance is quite consistent over all three test videos. Some visual comparisons for two test videos can be seen from Fig. 7, where we can see obvious advantages of our method even under occlusion. The comparison above shows the advantage of using the combined spatio-temporal priors over one with the spatial prior alone.

Table 1. The comparison of localization error (in pixel).

Methods	Head	Torso	R_arm	L_arm	R_leg	L_leg
1-fan [3]	5.3	6.8	12.2	11.1	12.7	12.8
hybrid [2]	5.9	6.0	9.8	8.6	4.8	5.6
vr03_7	0.6	0.9	1.3	1.8	1.3	1.0
vr08_7	0.6	1.8	0.9	4.8	1.7	1.1
vr21_7	0.9	1.0	1.5	7.4	2.7	2.7
Average	0.7	1.3	1.2	4.7	1.9	1.6

7.2. Special Case Handling

One major advantage of online learning is the ability to handle unseen motion patterns and occlusion. For example, Sequence vr14_7 shows an abnormal walking pattern that is hard to cope with for a tracker that relies on off-line learning. Moreover, one arm is occluded most of time during the walking cycle. The two competing algorithms fails in this case since the spatial prior learned off line is not able to deal with this abnormality. However, the proposed algorithm can accurately detect all visible body parts, regardless of the unusual motion pattern and one occluded arm. Some tracking results are shown in Fig. 8. It is shown that the proposed tracker can effectively localize the arm that deviates from its normal motion pattern, proving the usefulness of online learning. Although the majority of one arm is occluded in most frames, the tracking results of other body parts are not affected, showing the effectiveness of occlusion handling.

Figure 8. Tracking results for an abnormal activity.

7.3. Part Segmentation

The proposed algorithm can also online learn part-based appearance models from frame to frame during tracking. Part detection in this work is similar to the one used in [2]. After we localize the whole body (we use the Torso as the reference part) in the current frame, we can localize and segment each body part correspondingly. The segmented body parts can be used to speed-up part detection by providing an object template that can be updated sequentially by the

tracker. Some examples of online learned part appearances are shown in Fig. 9. Although each body parts exhibit significant shape/color variability among four test sequences, the segmentation results are quite accurate and robust.

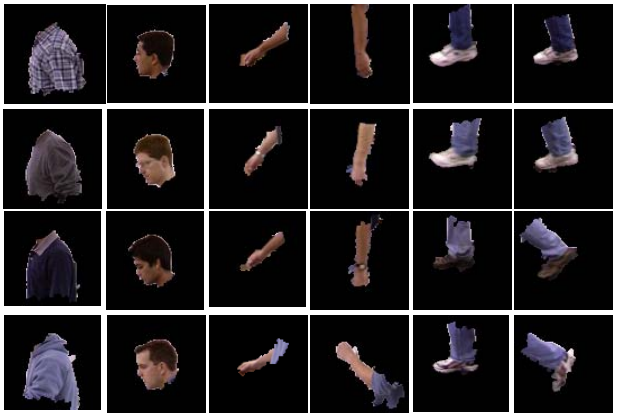

Figure 9. Online learned part appearance.

7.4. Discussion

Compared with the methods using holistic appearance models (e.g., [12]), this work is focused on articulated body tracking that is able to detect, track and segment body parts. Moreover, the major advantage of using part-based spatial and temporal priors is the ability of occlusion handling during tracking. Also, the part-based representation makes model learning more flexible that can involve both online and offline learning and be extended to handle some unseen activities. One may wonder how about the tracking results of using the temporal prior only. It was shown in our experiments, when the background is clean and no occlusion, the temporal prior alone could be sufficient given reasonable part-based appearance models. However, when the background is cluttered (with many false alarms) or occlusion occurs, the contribution from the spatial prior cannot be neglected. Other temporal prior models could be possible, such as Kalman filters. However, the major challenge will be the high dimension of the state space considering the number of body parts and the possible configuration of each body part, while BC-GPLVM is able to provide a LD latent space for effective state prediction via extrapolation.

8. Conclusion

We have proposed a new algorithm for articulated human tracking that combines both the spatial and temporal priors in an online learning framework. Compared with prior efforts, we want to fully take advantage of both the spatial and temporal priors in a balanced way in order to optimize the tracking performance. Although there might be certain redundancy between the two priors, the synergistic use of them greatly enhances the robustness and flexibility of the tracker, especially in a challenging environment with complex background or occlusion. The online learning mechanism makes the proposed algorithm effective to track subjects with significant appearance and motion variability. All of these makes our algorithm a promising tool to support video-based human motion analysis in a general setting.

Acknowledgements

This work is supported in part by the National Science Foundation (NSF) under Grant IIS-0347613, and an Oklahoma NASA EPSCoR Research Initiation Grant in 2009, and an OHRS award (HR09-030) from the Oklahoma Center for the Advancement of Science and Technology (OCAST). The authors also thank the anonymous reviewers for their comments and suggestions.

References

[1] D. Burr and J. Ross. Vision: The world through picket fences. *Current Biology*, 14:381–382, 2004.

[2] C. Chen and G. Fan. Hybrid body represenation for integrated pose recognition, localization and segmentation. In *Proc. IEEE CVPR*, 2008.

[3] D. Crandall, P. Felzenszwalb, and D. Huttenlocher. Spatial priors for part-based recognition using statistical models. In *Proc. IEEE CVPR*, 2005.

[4] D.Kahneman, A. Terisman, and B. J. Gibbs. The reviewing of object files: object specific integration of information. *Cognitive Psychology*, 24:175–219, 1992.

[5] C. H. Ek, P. H. Torr, and N. D. Lawrence. Gaussian process latent variable models for human pose estimation. In *Machine Learning for Multimodal Interaction*, 2007.

[6] F. J. elleman and V. E. D.C. Distributed hierarchical processing in primate cerebral cortex. *Cerebral Cortex*, 1:1–47, 1991.

[7] 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.

[8] S. Hou, A. Galata, F. Caillette, N. Thacker, and P. Bromiley. Real-time body tracking using a gaussian process latent variable model. In *Proc. IEEE ICCV*, 2007.

[9] X. Lan and D. P. Huttenlocher. A unified spatio-temporal articulated model for tracking. *Proc. of IEEE CVPR*, 1:722–729, 2004.

[10] N. Lawrence. Probabilistic non-linear principal component analysis with gaussian process latent variable models. *Journal of Machine Learning Research*, 6:1783C1816, 2005.

[11] N. D. Lawrence and J. Quinonero-Candela. Local distance preservation in the gp-lvm through back constraints. In *Proceedings of the 23rd international conference on Machine learning*, pages 513 – 520, 2006.

[12] C.-S. Lee and A. Elgammal. Modeling view and posture manifolds for tracking. In *Proc. IEEE ICCV*, 2007.

[13] D. Ramanan, D. A. Forsyth, and A. Zisserman. Tracking people by learning their appearance. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 29:65–81, 2007.

[14] L. Raskin, E. Rivlin, and M. Rudzsky. Using gaussian process annealing particle filter for 3d human tracking. *EURASIP Journal on Advances in Signal Processing*, 2008.

[15] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isardy. Tracking loose-limbed people. In *Proc. IEEE CVPR*, 2004.

[16] L. G. Ungerleider and M. Mishkin. *Two cortical visual systems*, pages 549 – 586. MIT Press, Cambridge, MA, USA, 1982.

[17] R. Urtasun and T. Darrell. Sparse probabilistic regression for activity-independent human pose inference. In *Proc. IEEE CVPR*, 2008.

[18] R. Urtasun, D. J. Fleet, and P. Fua. 3d people tracking with gaussian process dynamical models. In *Proc. IEEE CVPR*, 2006.

[19] J. M. Wang, D. J. Fleet, and A. Hertzmann. Gaussian process dynamical models for human motion. *IEEE Trans. Pattern Analysis and Machine Intelligence (PAMI)*, 30:283–298, 2008.