ARTICULATED HUMAN MOTION TRACKING WITH FOREGROUND LEARNING

Aichun Zhu[1]*, Hichem Snoussi*[1]*, Abel Cherouat*[2]

[1]ICD - LM2S - Université de Technologie de Troyes (UTT) - UMR STMR CNRS
[2]ICD - GAMMA3 - Université de Technologie de Troyes (UTT) - UMR STMR CNRS
12 rue Marie Curie - CS 42060 - 10004 Troyes cedex - France
E-mail :{aichun.zhu, hichem.snoussi, abel.cherouat}@utt.fr

ABSTRACT

Tracking the articulated human body is a challenging computer vision problem because of changes in body poses and their appearance. Pictorial structure (PS) models are widely used in 2D human pose estimation. In this work, we extend the PS models for robust 3D pose estimation, which includes two stages: multi-view human body parts detection by foreground learning and pose states updating by annealed particle filter (APF) and detection. Moreover, the image dataset F-PARSE was built for foreground training and flexible mixture of parts (FMP) model was used for foreground learning. Experimental results demonstrate the effectiveness of our foreground learning-based method.

Index Terms— Annealed particle filter, human motion tracking, foreground learning

1. INTRODUCTION

Articulated human motion tracking is a fundamental task in computer vision, which is widely used in medical science fields, visual surveillance and driving assistance systems, human computer interaction, etc. Markerless full-body tracking from images is a challenging work, and variety research work have been conducted in this area. There are two main methods for human motion tracking. The first method is tracking by detection, which is based on a single frame. The second one is the particle filter by using transition state models.

For human motion detection and recognition, the pictorial structure (PS) model is an influential approach [1], which decomposes the appearance of objects into local part templates with geometric constraints on pairs of parts. Felzenszwalb et al. [2] used PS model to realize object detection, which is called deformable part models. These models provide an elegant framework for object detection. Yang and Ramanan [3] proposed the flexible mixtures of parts (FMP) model, which is also based on PS framework. This model can be efficiently used in human body parts detection and pose estimation. As for tracking, Ramanan et al. [4] achieved human motion tracking by learning their appearance, which is still based on PS framework. As for 3D human pose estimation, multi-view techniques allow to achieve state of the art performance for more complex motions. Amin et al. [5] proposed multi-view PS models for robust 3D pose estimation.

Particle filtering [6] is one of the common approaches for human motion tracking, which used the pose in the current frame and a dynamic model to predict the next pose. Particle filter (PF) uses multiple predictions, obtained by drawing samples of pose and location prior, and then propagating them using the dynamic model by comparing them with the local image data and calculating the likelihood. The prior is typically quite diffused (because motion can be fast) but the likelihood function may be very peaky, containing multiple local maxima which are hard to account for in detail. Annealed particle filter (APF) [7] or local searches are the ways to tackle this problem. APF has been widely used for articulated human motion tracking due to its ability to precisely estimate the statistics of multi-modal and non-Gaussian processes. However, the performance of annealed particle filter drops when the frame rate is lower or the motion is moving fast.

This paper presents a new approach to track articulated human motion based on foreground learning by FMP model, which has shown strong ability to detect and estimate human motion from still images. In our work, we make use of the sequence to learn and subtract the background, and then jointly track and detect body parts in multiple views. Part models are trained based on F-PARSE dataset, which is the foreground of images in PARSE [8]. We evaluate our approach on the HumanEva-II dataset, which is a standard benchmark for 3D pose estimation. Finally, we empirically show the robustness of our approach under challenging conditions for human motion capture such as fast moving and self occlusion.

The rest of the paper is organized as follows: Section 2 describes particle filter for human motion tracking. Section 3 introduces foreground modeling by FMP model. Section 4 presents the proposed foreground learning based method for motion tracking. Implementation details are presented in Section 5. Finally, Section 6 draws the conclusion of this work.

This work is partially supported by China Scholarship Council of Chinese Government.

2. FILTERING

2.1. Particle filter

Particle filter algorithm was developed for tracking objects, using recursive Bayesian estimators derived from Monte Carlo sampling techniques. The algorithm aims at estimating the posterior density $p(x_t|y_{1:t})$, where $y_{1:t}$ notates the history of observation (x_t is a hidden state vector and y_t is a measurement at time t). The observation process is $p(y_t|x_t)$. The posterior density is represented by a set of weighted particles $\{(x_t^{(0)}, \pi_t^{(0)}) \cdots (x_t^{(N)}, \pi_t^{(N)})\}$, where $\pi_t^{(i)} \propto p(y_t|x_t^{(i)})$. The filtering distribution can be calculated using two steps.

Prediction step:

$$p(x_t|y_{1:t-1}) = \int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})dx_{t-1}. \quad (1)$$

Filtering step:

$$p(x_t|y_{1:t}) \propto p(y_t|x_t)p(x_t|y_{1:t-1}), \quad (2)$$

where $p(y_t|x_t)$ is the likelihood, and $p(x_t|y_{1:t-1})$ predicts the state at time t. Variations of PF: Sequential Importance Sampling (SIS) draws particles from a proposal distribution and then for each particle a proper weight is assigned as follows:

$$\pi_t^{(i)} \propto p(y_t|x_t^{(i)})p(x_t^{(i)}|x_{t-1}^{(i)})/q(x_t^{(i)}|x_{t-1}^{(i)}, y_t). \quad (3)$$

2.2. Annealed particle filter

APF has been used to track human motion because of its ability to tackle a function with several peaked local maxima. The main idea of APF is to utilize a series of weighting functions($w_0(y_t, x)$ to $w_M(y_t, x)$), where each $w_m(y_t, x)$ differs only slightly from $w_{m-1}(y_t, x)$. The weighting function $w_M(y_t, x))$ is designed to be very smoothed, representing the overall trend of the search space while $w_0(y_t, x)$ might be peaky. This is achieved by using $w_m(y_t, x) = (w_0(y_t, x))^{\beta_m}$, where $1 = \beta_0 > \cdots > \beta_M$ and $w_0(y_t, x)$ is equal to the original weighting function. Therefore, each annealing run includes M layers, and is started at layer M.

3. FOREGROUND MODELING

3.1. Basic pictorial structure model

Pictorial structure [1, 9] model for an object is given by a collection of parts with connections between certain pairs of parts. More specifically, for human body model, the parts can correspond to the head, torso, arms and legs of the human, as shown in figure 1. Pose parameters are optimized by maximizing the score function which is defined as follows,

$$S(I, L) = \Sigma_{i \in V} \alpha_i \cdot \phi(I, p_i) + \Sigma_{ij \in E} \beta_{ij} \cdot \psi(p_i, p_j), \quad (4)$$

where I denote the image, V is a set of nodes and p_i, p_j are locations of part i and j. α_i is unary template for part i, and

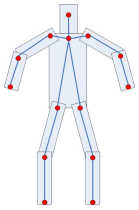

Fig. 1: Human body model based on pictorial structure: each node and link corresponds to a part and a physical connection between parts.

Fig. 2: Human body parts detection by flexible mixtures of parts model. First row shows the detection results are correct, while the second row shows FMP fails to detect body parts.

$\phi(I, p_i)$ is local image features at location p_i in image I; β_{ij} is pairwise springs between part i and part j, and $\psi(p_i, p_j) = [x_i - x_j, (x_i - x_j)^2, y_i - y_j, (y_i - y_j)^2]^T$ is the relative location between part i and part j.

3.2. Flexible mixtures of parts model

Flexible mixtures of parts model is also based on PS framework. As shown in the first row of figure 2, this model uses smaller body parts rather than the larger one, which is significantly faster than the original model. This section describes FMP model. Taking mixture of parts into account, the new score function can be defined as:

$$S(I, L, M) = \Sigma_{i \in V} \alpha_i^{m_i} \cdot \phi(I, p_i) + \Sigma_{ij \in E} \beta_{ij}^{m_i m_j} \cdot \psi(p_i, p_j) + S(M), \quad (5)$$

where m_i is the mixture of part i, $\alpha_i^{m_i}$ is unary template for part i with mixture m_i, and $\beta_{ij}^{m_i m_j}$ is pairwise springs between part i with mixture m_i and part j with mixture m_j. $S(M) = \Sigma_{ij \in E} b_{ij}^{m_i, m_j}$ is a sum of pairwise scores, and the pairwise parameter $b_{ij}^{m_i, m_j}$ favors particular co-occurences between part i with mixture m_i and part j with mixture m_j. E is a set of links each of which connect two parts. As shown in the second row of figure 2, we can note that this method can be confused by background. Contour-based features have

been proposed for articulated pose estimation [10], in an attempt to solve some of the background confusion situations. In our work, we make use of the sequence to learn and subtract the background, and then jointly track and detect body parts in multiple views.

4. TRACKING WITH FMP-APF

Based on only the annealing particle filter, one cannot efficiently track fast apparent motions due to low frame rates. On the other hand, FMP model cannot find some body parts due to the overlapping and occlusion. For these reasons, we combine these two methods together, and propose a foreground learning-based approach. Figure 3 is the illustration of this proposed FMP-APF scheme.

4.1. Modeling the body

As is common in the literature, we build the body model as a 3D kinematic chain with limbs, which consists of 15 segments: pelvis area, torso, head, upper and lower arms and legs, hands and feet. Our objective is to find the pose of the body over time, which is parametrized by a reduced set of 34 parameters comprising the global position and orientation of the pelvis and the relative joint angles between neighboring limbs. The shoulders, hips and thorax are modeled as ball and socket joints with 3 degrees of freedom, the clavicles are allowed 2 degrees of freedom, while the knees, ankles, elbows, wrists and head are assumed to be hinge joints requiring only one degree of freedom [11].

4.2. Likelihoods

For each particle in the posterior representation, the likelihood represents how well the projection of a given body pose state fits the observed images. Many image features could be used, including optical flow, color and adaptive appearance regions, however, the most common approaches are based on silhouette and edge information.

Edge-based log-likelihood function is estimated by projecting the pose into the edge map sparse points:

$$-\log p^e(y_t|x_t) \propto \frac{1}{k}\sum_{i=1}^{k}(1 - M_i^e(x_t, Y))^2, \qquad (6)$$

where Y is the image from which the pixel map is derived, and $M_i^e(x_t, Y)$ are the values of the edge pixel map at the K sampling points taken along the model's silhouette.

Silhouette-based log-likelihood function is estimated by projecting the pose into the foreground silhouette map sparse points:

$$-\log p^r(y_t|x_t) \propto \frac{1}{k}\sum_{i=1}^{k}(1 - M_i^r(x_t, Y))^2, \qquad (7)$$

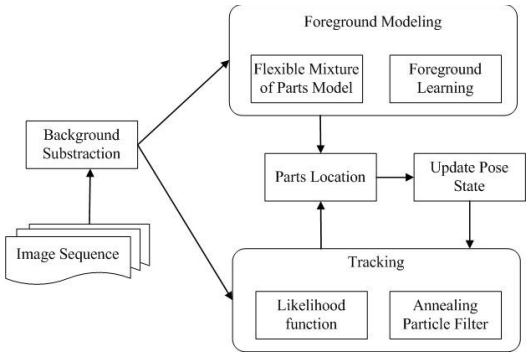

Fig. 3: Illustration of the proposed method.

where $M_i^r(x_t, Y)$ are the values of the foreground silhouette pixel map at the K sampling points taken from the interior of the model.

4.3. Detection by FMP in multi-view scene

As discussed in Section 3, FMP fails to detect body parts, because of overlapping and occlusion. Multiple views have a powerful ability to solve these problems by combining the detection in each view. So, this paper extends FMP to the multi-view case:

$$S(I, P, M, K) = \Sigma_{i\in V}\alpha_i^{m_i} \cdot \phi(I_k, p_{i,k}) + \Sigma_{ij\in E}\beta_{ij}^{m_i m_j} \cdot \psi(p_{i,k}, p_{j,k}) + S(M), \quad (8)$$

where I_k denotes the image I in view k, $p_{i,k}$ is the location of part i in view k, and $S(M)$ is a sum of pairwise scores. Let (n, m) denotes a pair of different views from K views. So $p_{i,n}$ and $p_{i,m}$ are the locations of part i in views n, m, which are calculated by eq.8. However, sometimes the position p_i is not enough accurate. Epipolar constraint is used between two views to remove false measurements and to achieve more accurate localization. The fundamental matrix F is the representation of epipolar geometry and the epipolar constraint is represented by $p_{i,n}^T F p_{i,m} = 0$. If points $p_{i,n}$ and $p_{i,m}$ are coherent, the $p_{i,n}$ lies on the epipolar line $l = Fp_{i,m}$. In this case, the 3D position q_i of part i can be computed by the back-projection of $p_{i,n}$ and $p_{i,m}$ as follows,

$$\begin{cases} L_{i,n}(\lambda) = P^+ p_{i,n} + \lambda C, \\ L_{i,m}(\lambda) = P^+ p_{i,m} + \lambda C, \end{cases} \qquad (9)$$

where $L_{i,n}, L_{i,m}$ are two rays, P^+ is the pseudo-inverse of camera matrix P, and C is the camera center. The intersection of the two rays $L_{i,n}, L_{i,m}$ is the 3D position q_i. From all possible (n, m) of the K views, it at least one pair is coherent, then the 3D position is retained and we consider the next body part. Otherwise, an update of the previous of 3D position is performed by APF as detailed in next subsection.

4.4. Update the state with APF

As discussed above, some body parts don't have any multi-view correspondence by FMP. To solve this, we introduce APF in FMP framework to realize robust tracking for all body parts. From APF, the optimal configuration have been computed from the particle set at the bottom layer using:

$$x_{t-1} = \Sigma_{j=1}^{N_p} \pi_{t-1,0}^{(j)} x_{t-1,0}^{(j)}, \tag{10}$$

where N_p is the number of particles. Let $x_{t-1} = (X_{t-1,1}, X_{t-1,2} \cdots X_{t-1,S})$, $X_{t-1,i}$ is the parameter vector for part i at time $t-1$, and S is the number of body parts. As discussed in Section 2, after the sample is drawn, the state estimation for each particle becomes:

$$p(x_t|x_{t-1}, y_t) \propto p(y_t|x_t)p(x_t|x_{t-1}). \tag{11}$$

APF is not appropriate for estimating high dimensional state parameters, especially for the state parameters of fast move body parts (arms and legs). The main idea of this paper is to use the detection of body parts to infer a subset of the state parameters. Suppose that the state vector x_t can be decomposed into $(x_{t,1}, x_{t,2})$, where $x_{t,1}$ is to be computed by APF, while the state parameters $x_{t,2}$ have already been computed by multi-view FMP. Therefore, the state estimation for each particle can be rewritten as:

$$p(x_{t,1}|x_{t-1,1}, x_{t,2}, y_t) \propto p(y_t|x_{t,1}, x_{t,2})p(x_{t,1}|x_{t-1,1}, x_{t,2}), \tag{12}$$

the above expression combines tracking and detection to perform automatic recovering from body-parts tracking failures. As represented by the term $p(x_{t,1}|x_{t-1,1}, x_{t,2})$, which is used to estimate the state $x_{t,1}$ based on the parameter $x_{t-1,1}$ and $x_{t,2}$. After all particles are computed, the optimal configuration have been computed at the bottom layer as follows:

$$x_{t,1} = \Sigma_{j=1}^{N_p} \pi_{t,1,0}^{(j)} x_{t,1,0}^{(j)}, \tag{13}$$

so the new state vector x_t is also computed (see Algorithm 1).

Algorithm 1 APF-FMP.

Input: Images $I_{t,k}$ from views k at time t ($k = 1 \cdots K$), state vector x_{t-1} at time $t-1$.
for $n = 1 \cdots K-1$
  for $m = n+1 \cdots K$
    Compute $F_{n,m}$ between views n,m
  end
end
% Applying multi-view FMP
for $i = 1 \cdots S$ % body part i
  for $n = 1 \cdots K-1$
    for $m = n+1 \cdots K$
      if $p_{i,n}'^T F_{n,m} p_{i,m}' == 0$
        Compute rays $L_{i,n}(\lambda) = P^+ p_{i,n}' + \lambda C$
        $L_{i,m}(\lambda) = P^+ p_{i,m}' + \lambda C$
        Compute q_i by the intersection of $L_{i,n}, L_{i,m}$
        Update the parameter vector $X_{t,i}$ with q_i
        $x_{t,2}(i) = X_{t,i}$
      end
    end
  end
end
Set the state vector $x_t = (x_{t,1}, x_{t,2})$
% $x_{t,1}$: non-matching with FMP
% $x_{t,2}$: matching with FMP
Compute $p(x_{t,1}|x_{t-1,1}, x_{t,2}, y_t)$ for each particle
Compute $x_{t,1} = \Sigma_{j=1}^{N_p} \pi_{t,1,0}^{(j)} x_{t,1,0}^{(j)}$
return x_t

5. IMPLEMENTATION DETAILS

We conducted a series of experiments to measure the effectiveness of our proposed models in real multi-view 3D settings on a variety of sequences from the HumanEva-II dataset.

Datasets. HumanEva [7] is a standard benchmark for 3D human pose estimation in the laboratory setting, which allow quantitative evaluation of performance. The dataset consists of HumanEva-I and HumanEva-II by a set of multi-view sequences. We utilize sequences of walking, jogging and balancing from HumanEva-II for our experiments. As for foreground training, we have built a dataset, which is the foreground from images in PARSE dataset, and the annotation of human body parts is not changed. We called this dataset as F-PARSE. There are some images from F-PARSE, as show in figure 4.

Evaluation of our approach. We evaluate the performance of our approach on HumanEva-II by the measure proposed in [7], which computes the 3D errors in millimeters of the locations of the joints and end points of the limbs between 15 virtual markers on the body and detection results. Then we compare performance against the baseline algorithm based on the methods of Deutscher and Reid [12], which have the same likelihoods and the same number of samples. Balan et al. [13] report APF with edge-based and silhouette-based likelihood function with 5 layers (200 particles per layer). The errors of their work reach 263±60mm for tracking the first 150 frames

Fig. 4: Images from F-PARSE dataset.

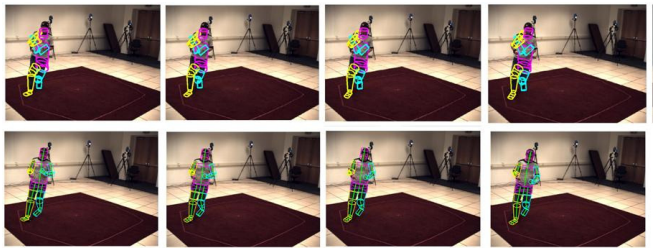

Fig. 5: Comparison of motion detection. First row shows motion detection by baseline algorithm. Second row shows the detection by FMP-APF.

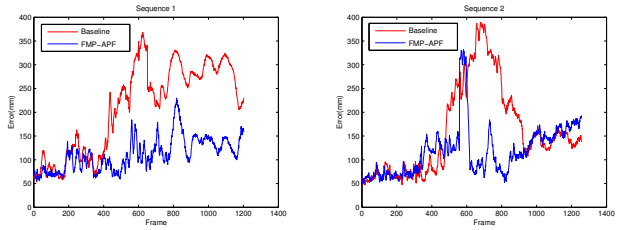

Fig. 6: Comparison of errors. The first 400 frames are for walking, and 401-700 frames are for jogging, and the rest for balancing. Left: 3D errors for the first subject by baseline algorithm and FMP-APF. Right: 3D errors for the second subject.

of the sequence. We applied standard particle filtering with foreground learning and compared our proposed method with baseline in figure 5 by computing the 3D errors in millimeters of HumanEva II. The performance is clearly improved by our method, especially for jogging, as shown in figure 6.

6. CONCLUSION

In this paper, we proposed a new framework for human body parts tracking, which is based on flexible mixture of parts model and annealing particle filter. FMP model is used for foreground learning in multiple views, and APF is used for tracking body parts. Then jointly track and detect body parts by estimating and updating the pose state. Experimental results have shown that the proposed method can efficiently track fast change motions.

In future research, we will do some improvements in our system. Firstly, we will try to optimize our tracking system, which is still not enough for robust tracking. Secondly, we will focus on online machine learning to improve the performance of detection.

REFERENCES

[1] P. Felzenszwalb and D. Huttenlocher, "Pictorial structures for object recognition," *International Journal of Computer Vision*, vol. 61, no. 1, pp. 55–79, 2005.

[2] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[3] Y. Yang and D. Ramanan, "Articulated human detection with flexible mixtures-of-parts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, pp. 1–14, 2012.

[4] D. Ramanan, D. Forsyth, and A. Zisserman, "Tracking people by learning their appearance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 65–81, 2007.

[5] S. Amin, M. Andriluka, M. Rohrbach, and B. Schiele, "Multi-view pictorial structures for 3d human pose estimation," in *Proceedings of British Machine Vision Conference (BMVC)*, pp. 1–12, 2013.

[6] M. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

[7] L. Sigal, A. Balan, and M. Black, "Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion," *International Journal of Computer Vision*, vol. 87, no. 1-2, pp. 4–27, 2010.

[8] D. Ramanan, "Learning to parse images of articulated bodies," in *Advances in Neural Information Processing Systems*, pp. 1129–1136, MIT Press, 2007.

[9] H. Bhaskar, L. Mihaylova, and S. Maskell, "Articulated human body parts detection based on cluster background subtraction and foreground matching.," *Neurocomputing*, vol. 100, pp. 58–73, 2013.

[10] N. Ukita, "Articulated pose estimation with parts connectivity using discriminative local oriented contours," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3154–3161, 2012.

[11] G. Taylor, L. Sigal, D. Fleet, and G. Hinton, "Dynamical binary latent variable models for 3d human pose tracking," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 631–638, 2010.

[12] J. Deutscher and I. Reid, "Articulated body motion capture by stochastic search," *International Journal of Computer Vision*, vol. 61, no. 2, pp. 185–205, 2005.

[13] A. Balan, L. Sigal, and M. Black, "A quantitative evaluation of video-based 3d person tracking," in *Proceedings of IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pp. 349–356, 2005.