2014 IEEE Winter Conference on Applications of Computer Vision

Robust Tracking of Articulated Human Movements through Component-based Multiple Instance Learning with Particle Filtering

Kyuseo Han Johnny Park Avinash C. Kak
School of Electrical and Computer Engineering, Purdue University
West Lafayette IN, USA
han66@purdue.edu jpark@purdue.edu kak@purdue.edu

Abstract

We present a robust approach for tracking human subjects as their limbs and torso are engaged in large articulated movements while the entire body is executing a large translational motion with respect to the pointing angle of the camera. While the articulated movements can be handled by the recently proposed Component-Based Multiple Instance Learning (CMIL) tracker, the large translational motions by the target require that we also use a motion prediction framework to more accurately estimate the most probable positions of the target in the next frame of a video sequence. In the work we report here, this prediction is carried out with a particle filter. This coupling between CMIL based tracking and particle filtering yields a much more accurate estimate of candidate positions of the target in the next frame given the position of the target in the current frame. We validate this new approach by demonstrating results on videos of human subjects that are simultaneously executing large articulated movements with their limbs and torso while the subjects themselves are in some translational motions with respect to the pointing angle of the camera.

1. Introduction

Tracking humans is important in several computer vision applications, e.g., in surveillance, pedestrian safety, sports broadcasting, behavior analysis, and so on. Tracking results that one sees most often in the literature usually involve only simple motions such as walking and running. Since these algorithms depend straightforwardly on projecting — in some cases after taking into account the motions estimated for the subject — a bounding box at the location of the subject in the current frame into a bounding box in the next frame, they fail for obvious reasons when the human subjects are executing large articulated motions [3].

There do exist tracking approaches that come under the label "Tracking by Detection" that can be expected to work more robustly in the presence of large articulated motions. In general, their performances depend on the accuracy with which the various components of the articulated object being tracked can be detected in a frame under the typical constraints of real-time processing of a video stream. The accuracy versus time tradeoff in these algorithms can be improved by organizing the components in some manner in one frame for the purpose of searching for their correspondents in the next frame. Approaches based on Multiple Instance Learning (MIL) [2, 9] are one way to solve this problem of how to lend some organization to this frame-to-frame search for the different components of an articulated object.

The MIL framework for object tracking [1, 4] is based on using positive and negative bags of instances for learning a binary classifier so that the classifier can detect the object being tracked among the various candidate bounding boxes in the next frame. The positive instances are constructed by translational variations of the boundaries of the most probable bounding box in the next frame. And, the negative instances are constructed from the image pixels far from the most probable bounding box. (The notion of MIL in machine learning was first advanced in [2, 9].) The work reported in [5] presents a variation on the basic MIL tracker to make it more suitable for tracking large *in-place* articulated motions. This approach is known as the Component-based MIL (CMIL) because a key step in the algorithm is the application of automatic segmentation to the positive and negative instances. The segmentation of positive and negative instances yields components that lend themselves more easily to the delineation of the pixel blobs in the next frame.

Whereas the CMIL approach of [5] works well for large but in-place articulated motions, it breaks down for obvious reasons when the human subject is engaged in large translational motions while his/her limbs and torso are engaged in large articulated movements. Both the originally-proposed MIL based tracking [1] and the CMIL based tracking [5] assume that the frame-to-frame variations in the center of

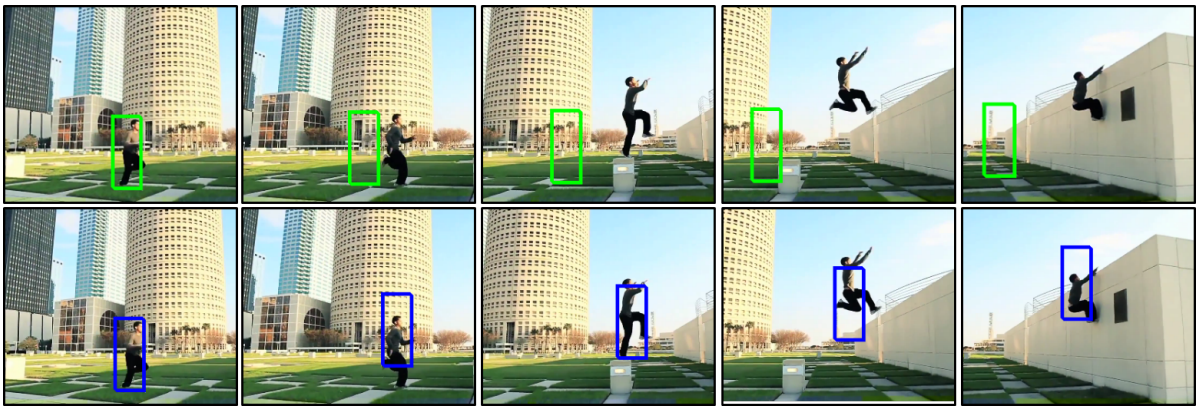

Figure 1. Note that the human subject is engaged in both the articulated movements made by his entire body while he is also executing a large translational movement with respect to the pointing angle of the camera. The sequence of images in the top row is for the case when tracking is attempted with just the CMIL based tracker. And the sequence of images in the bottom row is for the case when we combine particle-filter based motion prediction with CMIL based tracking.

mass of the target are sufficiently small so that the most probable bounding box in the next frame can be located simply by projecting the most probable bounding box in the current frame and searching in the vicinity of the projected bounding box. However, this assumption is violated if the human subject is also executing large translational motions with respect to the camera pointing angle. Obviously then, when a human subject is executing large translational motions while engaged in articulated movements of his/her limbs, head, and torso, we must combine the CMIL algorithm with a motion prediction framework. Our contribution is to demonstrate how to combine the CMIL-based tracking with the motion prediction framework, a particle filter. The top row of images in Fig. 1 visually illustrates how easy it is to lose a track when a motion prediction framework is not used to augment a CMIL based tracker. The entire body of the human subject is undergoing large articulated movements while the center of the blob of the pixels occupied by him is moving rapidly with respect to the pointing angle of the camera. However, when we include a particle-filter based motion prediction in the tracking algorithm, the CMIL tracker produces the excellent tracking results shown in the bottom row of images in the same figure. The remainder of this paper is organized as follows: In Section 2, we review the previously related work. Section 3 presents brief overviews of CMIL-based tracking and particle filter based tracking. In Section 4, how CMIL can be combined with a particle-filter based prediction framework. We present quantitative and qualitative experimental results produced by this combination tracker in Section 5. Finally, we conclude in Section 6.

2. Related Works

Since the inspiration for the work reported in this paper has come from how various researchers have combined novel approaches to target modeling with the Bayesian logic of particle filtering, we provide a brief review of such literature here. Another major source of our inspiration was how researchers have combined the particle-filter based tracking with binary classification logic for more robust tracking. In the rest of this section, we first quickly review the former and mention two specific contributions in which the human body was represented by a set of blobs. Subsequently, we describe the previous work on combing particle filtering with binary classification logic that is more along the lines of our own contribution. For human tracking, Lee et al. [7] have shown how a particle filter can be combined with a parts-based human tracker in which the human body is modeled as a set of parts and each part considered a particle in the particle filter. Along the same lines, Isard and MacCormick [10] have used a particle filter framework for multi-blob tracking with the blob likelihood representing the frame-to-frame location of the human body. With regard to combining particle filtering with binary classification logic for more robust tracking, Okuma et al. [13] have demonstrated an AdaBoost based approach that is used to combine the results obtained by a mixture of particle filters. The proposal distribution in their work is represented by a linear combination of the prior transition distribution and a Gaussian distribution corresponding to the detections at the output of the AdaBoost algorithm. Along similar lines, Li et al. [8] have proposed a framework for low frame rate video tracking that is based on using AdaBoost to combine three distinctive classification algorithms, LDA, offline AdaBoost, and an online-learned AdaBoost, in a three-stage cascade implementation. On the other hand, the contribution by Song et al. [14], deals specifically with the dynamic nature of the changes to the particles and also the number of particles as a target is being tracked. The particles in this approach are selected on the basis of the weights assigned

to them by an SVM classifier. As the target is being tracked, the particles that become invisible for one reason or another are assigned uniform weights. Closer to home, there is the work reported in [11, 15] in which an appearance based model of the target is learned and and continually updated with MIL and the model then tracked with a particle-filter based tracker. The MIL framework in these contributions helps the tracker cope with noisy nature of on-line learning.

In relation to the contributions mentioned above, our goal in this paper is to demonstrate that when we combine the CMIL-based tracking with motion prediction made possible by particle filtering, we get a truly robust tracker that can deal with the body articulation by a human target as the target is moving across the field of view of a camera.

3. Brief Reviews of MIL, CMIL, and Particle Filtering for Tracking

3.1. MIL and CMIL Based Tracking

The main idea of MIL is to learn the best class labels for a set of data instances from what are known as the positive and the negative *bags* of such instances. A positive bag must contain at least one truly positive instance and all of the instances in a negative bag must be negative instances. The advantage of the MIL approach is that it is more accommodating of the errors made in labeling positive instances during the learning process. This gives an MIL-based object tracker, as, for example, originally formulated by Babenko et al. [1], the freedom to make errors when declaring certain blobs in the next frame as positive instances of the most probable blob in the current frame. All that is needed is that a positive bag for localizing the target in the next frame contain at least one truly positive instance. This can be ensured by straightforwardly projecting the most probable bounding box in the current frame into the next frame and creating several candidate bounding boxes by shifting this projected bounding box around. Assuming that the frame-to-frame motion of the target is small, we can be reasonably certain that the set of bounding boxes thus created in the next frame will contain at least one truly positive instance of the target. To ensure that all the instances in a negative bag are negative, all we have to do is to choose these instances relatively far from where the target being tracked is likely to be.

A fundamental step in an MIL-based tracker is the transfer of probabilities to one or more candidates for positive instances in the next frame from one or more positive positive instances in the current frame. In the past, this has been carried out on the basis of pixel brightness similarities between the bounding boxes. However, this logic for establishing similarities breaks down when a human subject is executing large articulated motions. In such cases, estimating similarity probabilities on the basis of the similarity of pixel brightness levels — especially when such calculations

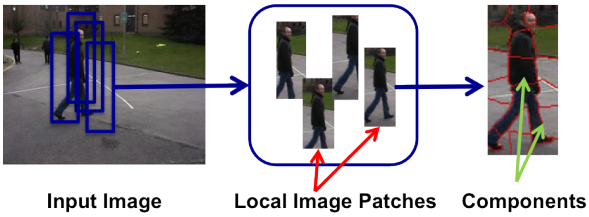

Figure 2. Example of local image patches and components

involve all of the pixels inside the bounding boxes — just does not work. In CMIL [5], this problem was taken care of by applying automatic segmentation to the bounding boxes in the current frame to yield what the authors of [5] refer to as the *components*, as shown in Fig. 2. (Hence the name CMIL for the algorithm in [5].) As the authors of [5] have argued, since each segmented component is likely to have a fairly uniform color, establishing correspondences between the positive instances in the current frame and the next is likely to be more accurate than when the same calculations are carried out with all of the pixels inside the bounding boxes. This allowed the authors of [5] to demonstrate tracking results when the human subjects were executing large articulated motions. For the underlying MIL logic, according to the authors of [5], the CMIL was implemented using the MILBoost algorithm of Viola et al. [16]. The label assigned assigned to a candidate positive instance in their implementation was a "noisy-or" of the labels of the components contained therein.

3.2. Particle Filter Based Tracking

For a quick review of the formulas that go into a prediction framework based on particle filtering, consider a sequence of the state vectors $\{\mathbf{x}_t \mid t \in \mathbb{N}\}$ and another sequence of the observation vectors $\{\mathbf{z}_t \mid t \in \mathbb{N}\}$ that we may associate with a target in motion. We assume that the time evolution of the state vectors is described by a possibly nonlinear function f_t as shown below (this being referred to as the *process model*):

$$\mathbf{X}_t = f_t(\mathbf{X}_{t-1}, \mathbf{V}_t). \tag{1}$$

We may also associate an observation model with the sequence of state vectors:

$$\mathbf{Z}_t = g_t(\mathbf{X}_t, \mathbf{W}_t), \tag{2}$$

where g_t is also possibly a nonlinear function. The $\mathbf{V}_t$ and $\mathbf{W}_t$ are the white noise and the observation noise.

From a Bayesian perspective, the tracking problem is to recursively compute a Bayesian estimate for $\mathbf{x_t}$ given the observations $\mathbf{z}_{1:t}$ up to time t. The prediction problem is to construct a Bayesian estimate for $\mathbf{x_{t+1}}$ given the observations $\mathbf{z}_{1:t}$ up to time t. Thus it is required to construct the posteriori state probabilities for either $p(\mathbf{x}_t|\mathbf{z}_{1:t})$ or $p(\mathbf{x}_{t+1}|\mathbf{z}_{1:t})$ given all the causal observations.

In particle filter, in general, the posteriori probabilities for $p(\mathbf{x}_t|\mathbf{z}_{1:t})$ are approximated by a set of N samples and their weights w, $\{\mathbf{x}_t^{(i)}, w_t^{(i)}\}_{i=1}^N$, as follows:

$$p(\mathbf{x}_t|\mathbf{z}_{1:t}) \approx \sum_{i=1}^{N} w_t^{(i)} \delta(\mathbf{x}_t - \mathbf{x}_t^{(i)}), \tag{3}$$

where $\delta(\cdot)$ is a Kronecker delta function. For object tracking in videos, one commonly uses the sequential importance resampling (SIR) particle filter for removing what is known as the degeneracy problem. In the prediction stage of an SIR particle filter, one starts by selecting new particles from the state transition probabilities $p(x_t|x_{t-1})$ as follows:

$$\mathbf{x}_t^{(i)} \sim p(\mathbf{x}_t|\hat{\mathbf{x}}_{t-1}^{(i)}),\ i = 1, \cdots, N, \tag{4}$$

where $\hat{\mathbf{x}}_{t-1}^{(i)}$ is the sample at $t-1$ after resampling process. In the update stage, the posterior PDF $p(\mathbf{x}_t|\mathbf{z}_{1:t})$ is computed by updating the weight of each particle with the likelihoods as follows:

$$w_t^{(i)} \approx p(\mathbf{z}_t|\mathbf{x}_t^{(i)}). \tag{5}$$

4. CMIL Tracker with Particle Filter Based Motion Prediction

The proposed approach provides a framework that couples a CMIL tracker and a particle filter for robust tracking of articulated human movements while the entire body is executing a large translational motion. Figure 3 illustrates the difference between the conventional CMIL tracker [5] and the proposed CMIL tracker with particle filter based motion prediction. The conventional CMIL tracker uniformly samples positive image patches around the estimated position of the target in the previous frame. On the other hand, the proposed method adaptively samples positive image patches at the locations specified by the particle filter. The motion prediction and adaptive sampling made possible by particle filtering provide increased robustness against a large translational motion of the target.

The proposed method is summarized in Algorithm 1. Given, the target estimate from the previous frame, the particle filter first selects new particles using the state transition probability. The particle filter then computes the weight of each particle. For this purpose, we utilize two distance measures, d_B and d_{max}, which we will describe shortly. The posteriori probability is then approximated by a resampling step. Instead of uniformly sampling positive image patches as in the conventional MIL and CMIL, our method collects positive image patches at the locations of resampled particles and negative image patches elsewhere. Each of these positive and negative image patches is segmented into a set of components and fed into an online boosting algorithm to update the classifier $\mathbf{H}_t$ [5]. Given the updated classifier

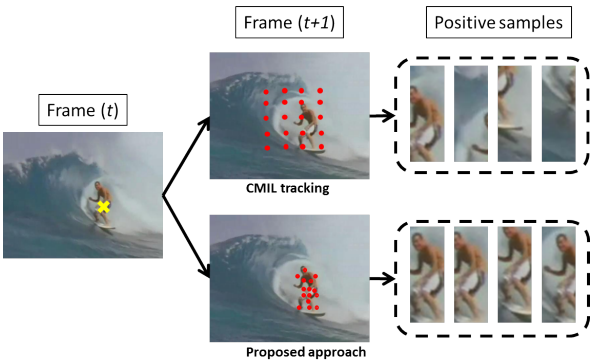

Figure 3. An illustration of the difference between the conventional CMIL tracker and the proposed CMIL tracker with particle filtering. The red dots indicate the sampling positions of a positive image patch at time $t+1$.

Algorithm 1 CMIL tracker with particle filtering

Input : $\{\hat{\mathbf{x}}_{t-1}^{(i)}, \hat{w}_{t-1}^{(i)}\}_{i=1}^N$,$\mathbf{H}_{t-1}$, and $\mathbf{s}_{t-1} = \mathbf{s}(\hat{\mathbf{x}}_{t-1})$

Output : $\{\hat{\mathbf{x}}_t^{(i)}, \hat{w}_t^{(i)}\}_{i=1}^N$, $\mathbf{H}_t$, and $\mathbf{s}_t = \mathbf{s}(\hat{\mathbf{x}}_t)$

Particle Filter Stage

- Predict Motion $\mathbf{x}_t^{(i)} \sim p(\mathbf{x}_t \mid \hat{\mathbf{x}}_{t-1}^{(i)})$
- Compute weight $w_t^{(i)} \approx p(\mathbf{z}_t \mid \mathbf{x}_t^{(i)})$
- Resample particles

$$\{\mathbf{x}_t^{(i)}, w_t^{(i)}\}_{i=1}^N \Rightarrow \{\hat{\mathbf{x}}_t^{(i)}, \hat{w}_t^{(i)}\}_{i=1}^N$$

CMIL Stage

- Extract image patch at each $\hat{\mathbf{x}}_t^{(i)}$
- Run CMIL online boost for updating
- Compute the updated target feature $\mathbf{s}(\hat{\mathbf{x}}_t)$

$\mathbf{H}_t$, we then compute the feature vector $\mathbf{s}$ of the image patch that corresponds to the estimated position of the target. We mentioned earlier that we utilize two distance measures for the purpose of computing the weight of each particle. Before we define these two distance measures, we first need to introduce the feature vector $\mathbf{s}$ of an image patch. Recall that each of the positive and negative image patches is segmented into the set of components. The current classifier $\mathbf{H}_t$, when applied to each of these components, produces a *confidence score* of whether that component belongs to a positive instance. Let $\boldsymbol{\nu} = \{\nu^{(\ell)}|\ell = 1, \cdots, L\}$ be a set of confidence scores where $\nu^{(\ell)}$ is the output of the classifier $\mathbf{H}_t$ applied to the ℓ-th component in the image patch.

Based on the set of confidence scores, the feature vector $\mathbf{s}$ is composed of two parts: the first part is the histogram of the confidence score set $\boldsymbol{\nu}$, and the second part is simply the maximum value in the confidence score set. Denoting the histogram as $\mathbf{r}$ and $\nu_{max} \equiv \max(\boldsymbol{\nu})$, the feature vector of an image patch at $\mathbf{x}_t^{(i)}$ is defined as:

$$\mathbf{s}(\mathbf{x}_t^{(i)}) = [\mathbf{r}(\mathbf{x}_t^{(i)}), \nu_{max}(\mathbf{x}_t^{(i)})]. \tag{6}$$

Now that we have defined the feature vector of an image patch, we can describe how the weight of each particle is computed. Note that computing the weight of each particle involves computing the similarity between the feature vector of the image patch at the current estimate of the target (i.e., $\mathbf{s}(\hat{\mathbf{x}}_t)$) and the feature vector at each of the particle samples (i.e., $\mathbf{s}(\mathbf{x}_{t+1}^{(i)})$). We use the Bhattacharyya distance [12] as the first distance measure for comparing the two histograms. The Bhattacharyya coefficient measures the degree of overlap between two different discrete distributions, p and q:

$$\rho(p,q) = \sum_{i=1}^{N} \sqrt{p(i)q(i)}, \tag{7}$$

and the Bhattacharyya distance is defined as

$$d_B(p,q) = \sqrt{1 - \rho(p,q)}. \tag{8}$$

As demonstrated in [6], the maximum confidence score in a positive bag in MIL can be used for estimating the probability of the positive bag . Put it in an equation, we have

$$\Pr(y = 1|\nu) \propto \frac{1}{e^{-\nu_{max}}}, \tag{9}$$

where $y \in \{-1, 1\}$ is the label of bag. The second distance measure between two feature vectors of image patches, therefore, is simply defined as

$$d_{max}(\nu_{max}^{(k)}, \nu_{max}^{(\ell)}) = \nu_{max}^{(k)} - \nu_{max}^{(\ell)}, \tag{10}$$

where $\nu_{max}^{(k)} = \nu_{max}(\mathbf{x}^{(k)})$. Finally, the weight of each particle is then given by

$$\begin{aligned} w_{t+1}^{(i)} &\approx p(\mathbf{z}_{t+1}|\mathbf{x}_{t+1}^{(i)}) \\ &\propto e^{-\gamma(\alpha d_B(\mathbf{r}(\hat{\mathbf{x}}_t), \mathbf{r}(\mathbf{x}_{t+1}^{(i)})) + (1-\alpha) d_{max}(\hat{\nu}_t, \nu_{t+1}^{(i)}))} \end{aligned}, \tag{11}$$

where $\hat{\nu}_t = \nu_{max}(\hat{\mathbf{x}}_t)$, $\nu_{t+1}^{(i)} = \nu_{max}(\mathbf{x}_{t+1}^{(i)})$, and γ and α are user-specified parameters.

5. Experiments

We tested our proposed tracker, a CMIL tracker coupled with a particle filter (CMIL-PF), on some challenging video sequences: *Gym*[1], *Skating*[2] , *StandToSit*[3], and *Caviar Wiggle*[4] sequences. These sequences include extensive articulated human movements such as bending, rolling, handstand, etc. These sequences also include large translational motions by the human subjects. We compare our algorithm to the online MIL tracker [1] and component-based MIL (CMIL) tracker [5]. For a fair comparison, all parameters involved in learning the classifiers are fixed in all the trackers and in all the test sequences. In our method, we have fixed the number of particles to be 50 and the noise variance of state transition to be 7 pixels in all test sequences. We manually specified the initial position of human to be tracked in the first frame. We have empirically chosen the parameters in Eq. (11) for computing the weight of each particle as $\gamma = 2.5$ and $\alpha = 0.3$ in all test sequences. We analyzed the performance of the trackers both qualitatively and quantitatively. Figure 4 shows the qualitative assessment of the trackers by displaying the bounding box at the position of the human subject estimated by each tracker in each frame. The center position of each rectangle is the best estimated position of the human subject in the image coordinate space. We also carried out two quantitative evaluations. First, we measured the pixel distance between the target position estimated by each tracker and the ground truth position that are manually collected in all of the test sequences. Figure 5 shows the pixel distance errors of the three trackers on each of the four test sequences. In most cases, the proposed CMIL-PF trackers has least pixel distance errors compared to the MIL and CMIL trackers. An exception occurs in the *Gym* sequence at around frame 20 where the CMIL tracker performs better than the CMIL-PF tracker. The exceptional pixel distance errors in these frames are caused by the fact that the human subject is located at the bottom area of the estimated bounding box from the CMIL-PF tracker; in other words, the human subject is not located at around the center of the bounding box. As shown in the second frame (the frame number 15) of the first row in Fig. 4 , the center of the bounding box from the CMIL-PF tracker(the blue rectangle) deviates from the center of the human being tracked while the bounding box enclosed the human subject. However, the center of the human subject is located at around the center of the bounding box from the CMIL tracker (the green rectangle). Recall that the target position is the center position of the bounding box, which is used for computing the pixel distance error. Table 1 shows the average pixel errors.

The second quantitative evaluation was carried out using precision plots, similar to the ones described in [1]. A precision plot shows the percentage of frames in which the

[1] http://www.youtube.com/watch?v=FJIrHgshB20
[2] http://cs.snu.ac.kr/research/~vtd
[3] http://vision.cs.uiuc.edu/projects/activity
[4] http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/

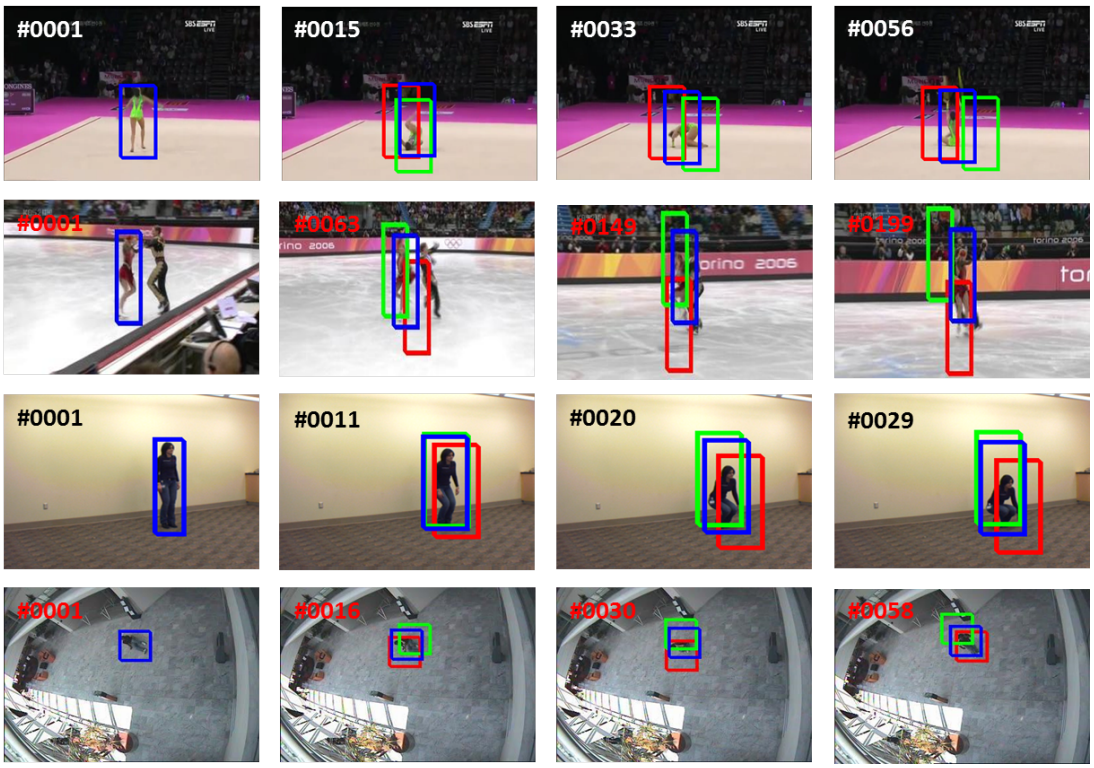

Figure 4. Tracking results shown are for the following four test sequences: *Gym*, *Skating*, *StandToSit*, and *Caviar Wiggle* from the top to bottom. The three trackers compared are: (1) MIL tracker (Red); (2) CMIL tracker (Green); and (3) the proposed CMIL-PF tracker (Blue).

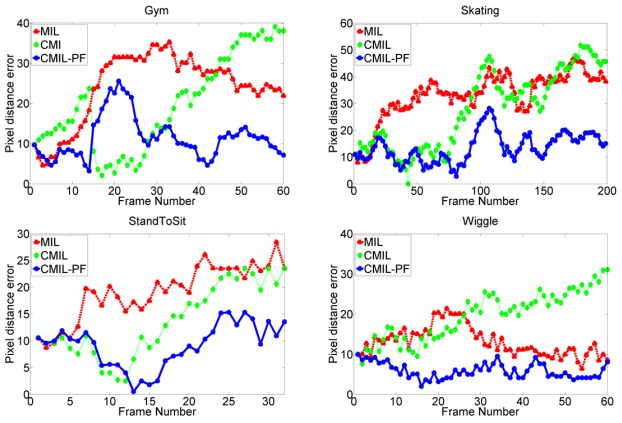

Figure 5. Pixel distance errors on four test sequences.

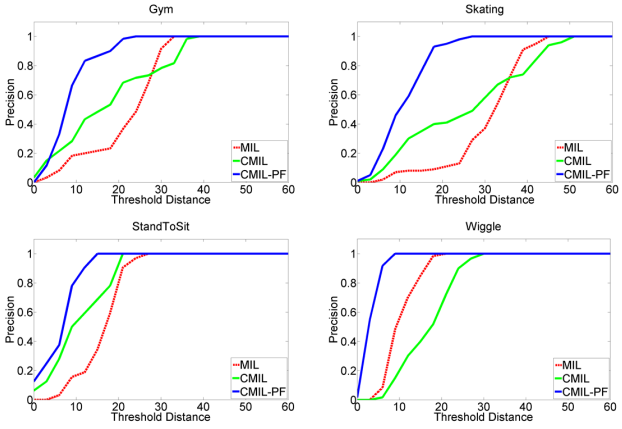

Figure 6. Precision plots for the three trackers on four test video sequences.

distance between the center position of estimated bounding box and the ground truth position was within some threshold distance. Furthermore, the precision is calculated from frames in which the detected human was located in the bounding box because the error distance is valid only when the detected human is inside the bounding box. Figure 6 shows the precision plots of the three trackers on each of the four test sequences. The precision plots clearly show that the proposed method increases the tracking accuracy of human subjects with large articulated movements and translational motions. In *Wiggle* sequence, for example, the CMIL-PF tracker was able to track the human subject with the accuracy of 10 pixels or less at all times. Note that only the qualitative result of *Parkour*[5] sequence is shown in Figure 1 because the MIL and CMIL trackers lost track within the first couple of frames.

[5]http://www.youtube.com/watch?v=XuiWzgdA6MA

	Gym	Skating	StandToSit	Wiggle
MIL track	24.11	33.15	19.19	13.11
CMIL track	19.76	27.94	13.62	19.6
CMIL-PF track	11.44	13.48	9.11	5.84

Table 1. Average pixel distance errors of compared trackers in each test sequence.

6. Conclusion

In this paper, we addressed the problem of tracking a human subject whose limbs, head, and torso are executing large articulated movements as the subject's body is moving rapidly across the image frame with respect to the pointing angle of the camera. It was already known in the research community that articulated movements by humans are best tracked with algorithms based on MIL (Multiple Instance Learning) since such algorithms are more forgiving of the errors made in declaring image blobs as being positive instances of the target being tracked. The next problem therefore was tracking the articulated movements while the human subject was also engaged in large translational motions with respect to the pointing angle of the camera. It is this problem that we have addressed and solved in this paper by combining the CMIL tracker with a particle filter. The motion prediction framework made possible by the particle filter makes it easier to identify in the next frame the best candidates for the positive instances given their location in the current frame.

We evaluated our tracking framework by comparing its performance against that of the other well-known tracking algorithms in video sequences containing extensive articulated and translational human movements. Our proposed tracker showed better tracking performance with respect to both the location errors and the precision of the track in all the test sequences.

References

[1] B. Babenko, M.-H. Yang, and S. Belongie. Robust object tracking with online multiple instance learning. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 33(8):1619–1632, 2010. 1, 3.1, 5, 5

[2] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Perez. Solving the multiple-instance problem with axis-paralle rectangles. *Artificial Intelligence*, 89:31–71, 1997. 1, 1

[3] P. Dollar, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(4):743–781, 2012. 1

[4] H. Grabner, M. Grabner, and H. Bischof. Real-time tracking via on-line boosting. In *Proc. Conf. British Machine Vision*, pages 47–56, 2006. 1

[5] K. Han, J. Park, and A. C. Kak. Tracking articulated human movements with a component based approach to boosted multiple instance learning. In *IEEE International Conference on Image Processsing*, 2013. 1, 3.1, 4, 4, 5

[6] M. Kim and F. De la Torre. Gaussian processes multiple-instance learning. In *International Conference on Machine Learning*, 2010. 4

[7] M. W. Lee, I. Cohen, and S. K. Jung. Particle filter with analytical inference for human body tracking. In *IEEE Workshop on Motion and Video Computing*, 2002. 2

[8] Y. Li, H. Ai, T. Yamashita, S. Lao, and M. Kawade. Tracking in low frame rate video: A cascade particle filter with discriminative observers of different lifespans. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1728–1740, 2008. 2

[9] O. Maron and T. Lozano-Perez. A framework for multiple-instance learning. In *Advances in Neural Information Processing Systems*, pages 570–576. MIT Press, 1998. 1, 1

[10] M.Isard and J. MacCormick. Bramble: A bayesian multiple-blob tracker. In *International Conference on Computer Vision*, 2001. 2

[11] Z. Ni, S. Sunderrajan, A. Rahimi, and B. Manjunath. Particle filter tracking with online mulitple instance learning. In *Internation Conference on Pattern Recognition*, 2010. 2

[12] K. Nummiaro, E. Koller-Meier, and L. V. Gool. An adaptive color-based particle filter. *Image and Vision Computing*, 21:99–110, 2002. Bhattacharyya coefficient and distance. 4

[13] K. Okuma, A. Taleghani, N. D. Freitas, J. J. Little, and D. G. Lowe. A boosted particle filter: Multitarget detection and tracking. In *the European Conference on Computer Vision*, 2004. 2

[14] C. Song, J. Son, S. Kwak, and B. Han. Dynamic resource allocation by ranking svm for particle filter tracking. In *British Machine Vision Conference*, 2011. 2

[15] Y. Song and Q. Li. Visual tracking based on mulitple instance learning particle filter. In *International Conference on Mechatronics and Automation*, 2011. 2

[16] P. Viola, J. C. Platt, and C. Zhang. Multiple instance boosting for object detection. In *Proc. Neural Information Processing System*, pages 1417–1426, 2005. 3.1