TRACKING ARTICULATED HUMAN MOVEMENTS WITH A COMPONENT BASED APPROACH TO BOOSTED MULTIPLE INSTANCE LEARNING

Kyuseo Han, Johnny Park, and Avinash C. Kak

School of Electrical and Computer Engineering, Purdue University

ABSTRACT

Our work is about a new class of object trackers that are based on a boosted Multiple Instance Learning (MIL) algorithm to track an object in a video sequence. We show how the scope of such trackers can be expanded to the tracking of articulated movements by humans that frequently result in large frame-to-frame variations in the appearance of what needs to be tracked. To deal with the problems caused by such variations, our paper presents a component based version of the boosted MIL algorithm. Components are the output of an image segmentation algorithm applied to the pixels in the bounding box encapsulating the object to be tracked. The components give the boosted MIL the additional degrees of freedom that it needs in order to deal with the large frame-to-frame variations associated with articulated movements.

Index Terms— Articulated human tracking, Multiple instance learning, Online boosting

1. INTRODUCTION

Tracking humans in motion is an important part of computer vision in applications that involve human subjects. Despite the fact that there now exist several algorithms for the same, the solutions obtained with the current algorithms are generally unsatisfactory in real-world applications [1].

For the problem of tracking humans when they are engaged in ordinary movements — such as walking — the solutions developed fall into two categories: those that are based on bounding boxes and those that do not. The bounding box based solutions basically track the image intensity distributions within a bounding box placed around the target in the first image of a sequence and then use either a Kalman filter [2, 3] or a particle filter [4, 5] to track the distribution. On the other hand, the algorithms that do not require a bounding box are based on using a probabilistic framework that can jointly carry out the segmentation that best describes the target in each frame and tracking [6, 7]. For tracking more complex human movements, researchers have proposed methods that use parts based model, with the parts standing for head, torso, arm, leg, belly, etc. [8, 9]. A major shortcoming of these methods is the need to identify the parts in the images. In some cases, one can get around this shortcoming by using a probabilistic framework [10, 11, 12]. However, even these approaches break down when extent of self-occlusion is significant. To deal with the problems caused by occlusion, researchers proposed tracking methods based on machine learning algorithms, such as a cascade of SVM classifiers[13], boosted cluster tree[14], etc.

The work we describe in this paper also deals with the problem of tracking human movements when the articulated motions span large variations in the configuration space. Our approach is based on discriminative learning called Multiple Instance Learning (MIL) [15, 16] that associates class labels with sets of instances (as opposed to single instances in the traditional learning algorithms). In recent years, MIL combined with existing classification strategies has been used to address problems in object detection and tracking [17, 18, 19]. Work has also been carried out in using MIL for online detection and tracking [20, 21]. In this work, the reference model used in the tracking process is dynamically updated from frame to frame.

Specific focus of our work described in this paper is on addressing the problem of tracking human movements that cause large frame-to-frame variations with the following steps: (1) Project the most probable bounding box in the previous frame into a candidate bounding box in the current frame; (2) Apply a strong classifier computed in the previous frame to different bounding boxes in the vicinity of the candidate box and select the most probable bounding box; (3) Construct randomly displaced versions of this most probable bounding box as positive examples; (4) Choose image patches from the current frame outside the region spanned by the positive examples as negative examples; (5) Apply an image segmentation algorithm to the positive and the negative examples; (6) Use the component based boosted MIL algorithm to assign probabilities to the different positive examples with regard to the similarity of their pixel distribution to the most probable bounding box in the current frame; and, finally, (7) Compute the strong classifier from the component based boosted MIL algorithm.

Whereas the overall logic of object tracking described above is not new, what is new is our version of the boosted MIL algorithm. In our approach, image patches in the positive and the negative examples are subject to automatic segmentation to yield what we call *components*. It is the

To appear in *International Conference on Image Processing(ICIP), Sep. 15-18, 2013, Melbourne, Australia*

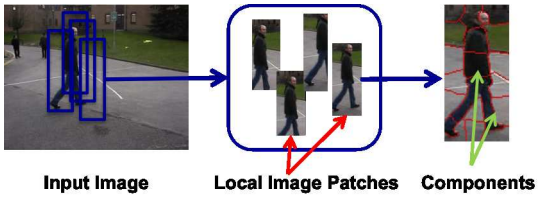


Fig. 1: Multiple image patches are first extracted from the input image. Automatic segmentation is then applied to each image patch to yield discriminative clusters, or components.

components that are subject to classification by the boosted MIL algorithm. The components give our approach extra degrees of freedom that are not possessed by the previous use for boosted MIL for object tracking. For articulated motions, we believe that these additional degrees of freedom can play a critical role in transferring the probabilities over the positive examples in the previous frame to those in the current frame. When articulated motions are involved, it is less likely that pixel brightness distributions would match well between two corresponding bounding boxes in two successive frames. However, if we first segment each of the bounding boxes into components based on, say, approximate uniformity of brightness levels, we are more likely to find component-to-component matches between the two corresponding bounding boxes. This then underlies the rationale behind our component based boosted MIL algorithm.

2. COMPONENT BASED MULTIPLE INSTANCE LEARNING APPLIED TO HUMAN TRACKING

In this section, we first provide a brief introduction to MIL and then show how this technique can be adapted for tracking humans with large articulated motions.

2.1. Multiple Instance Learning

MIL is a discriminative learning algorithm that is more accommodating of ambiguities in the training data than the conventional approaches. Given training samples (x_i, y_i), the goal is to estimate a function $f : \mathcal{X} \rightarrow \mathcal{Y}$, where $\mathcal{X} \in \mathbb{R}^d$ is the feature space with $x_i \in \mathcal{X}$, and $y_i \in \mathcal{Y}$ is the class label assigned to each data x_i. In MIL, the training data is not an individual sample but a set of the training samples called a bag and each training sample in the bag is called an instance. A bag X_i consists of a set of instances $X_i = \{x_{ij}\}_{j=1}^{N_i}$ and the label y_i of the bag X_i as assigned by $y_i = \max_j(y_{ij})$, where y_{ij} is the instance label, which is unknown during training stage. Each bag is either a positive bag or a negative bag. The positive bag $X^+ = \{x_j\}_{j=1}^{N^+}$ has to contain at least one positive sample x_j whereas the negative bag $X^- = \{x_j\}_{j=1}^{N^-}$ consists of all negative samples.

We need to estimate the function $f : \mathcal{X} \rightarrow \mathcal{Y}$ given the training samples $\{(X_i, y_i)\}_{i=1}^{N}$. Numerous algorithms for solving the MIL problem have been proposed [18, 22, 20] over the years. Of particular interest to us is the solution proposed by Viola et al. [19]. Their algorithm, called MILBoost, solves the MIL problem by training a boosted classifier that maximizes the log likelihood of the bags:

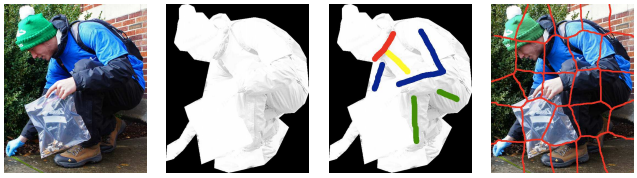
$$\log(\mathcal{L}(C)) = \sum_i \log(p(y_i \mid X_i)), \tag{1}$$

where y_{ij} is an output of classifier $y_{ij} = C(x_{ij})$. The likelihood of the bag $p(y_i \mid X_i)$ can be expressed in terms of the likelihood of each instance. The likelihood of a set of training bags is given by

$$\mathcal{L}(C) = \prod_i (p(y_i \mid X_i))^{y_i} (1 - p(y_i \mid X_i))^{(1-y_i)}, \tag{2}$$

where $y_i \in \{0, 1\}$ is a binary label. The likelihood $p(y_i \mid X_i)$ represents the connection with the bag X_i and its instances x_{ij} in terms of noisy-OR model

$$p(y_i \mid X_i) = 1 - \prod_{j=1}^{N_i} (1 - p(y_{ij} \mid x_{ij})). \tag{3}$$



(a) (b) (c) (d)

Fig. 2: Estimation of articulation of human pose. (a) an input image patch; (b) an obscure silhouette from foreground/background segmentation; (c) broken linkages and self occlusion in a possible joint model representation; (d) components in the image patch

2.2. Component Based Online Boosted MIL

A straightforward application of MILBoost fails to track human movements when large articulated motions are involved on account of excessively large variations in the appearance that needs to be tracked. To augment the logic of MILBoost in order to more explicitly take into account the articulation caused variations in the images, we now introduce a component based approach to the boosting of MIL in which the components provide us with the additional degrees of freedom to accommodate the image variations caused by articulations. By components we mean the blobs produced by an image segmentation algorithm when applied to the patch inside a bounding box. In this manner, our algorithm can simultaneously address the sort of appearance variations handled by MILBoost and the human articulations that we need for tracking people movements.

To explain how the components for the different image patches are used, in each frame we start with the extraction of

Algorithm 1 Component Based Boosted MIL

Input: Given N training samples (X_i, y_i) where image patches X_i and their correspondent label $y_i \in \{0, 1\}$,

1: Apply superpixel algorithm to obtain components $\{x_{ij}\}_{j=1}^{N_i}$ in i^{th} image patch X_i and set label $y_{ij}, \forall j, y_{ij} = y_i$
2: Extract features $\{f_{ijk}\}_{k=1}^{K}$ from each x_{ij}
3: Update all K weak classifiers in the pool with all data $\{x_{ij}, y_{ij}\}$
4: For each weak classifier,
5: Compute likelihood of components in each image patch
6: Compute likelihood of image patch by combining likelihood of components
7: Choose a classifier having maximum likelihood
8: Select the classifier and add to a strong classifier
9: Repeat until collecting M classifiers.

Output: $H(x) = \Sigma h_k(x)$

a number of image patches based on the current bounding box for the human being tracked. As in the regular MIL boosting, some of these patches are considered to be positive examples and the others negative examples. Each patch is segmented into components, as shown in Figure 1. While the appearance variations are addressed by collections of local image patches (i.e., positive instances), the human articulations are implicitly dealt with through the components. We can think of the components as implicitly representing the human parts. In this manner, our proposed boosted MIL approach can use the components in both the positive and the negative examples for accommodating human articulation changes during tracking.

To give the reader a sense of motion articulations as captured by the image components extracted from a patch, we show Figure 2. If we simply represent the human figure by a single patch containing a silhouette, as shown in Figure 2(b), the foreground/background segmentation shown in the figure that is used as a silhouette for tracking usually contains insufficient information for tracking the movements. Even a parts based approach may fail to track in such cases on account of the "broken linkages" shown in Figure 2(c) when we fit an articulated model to the silhouette. However, the components in the image patch, as shown in Figure 2(d), can implicitly represent human articulation with loosely-connected configurations in that each component contains partial or whole human parts.

We will now extend the MIL formalism of the previous section in order to incorporate the components in it. We now consider a bag X_i to consist of a set of sub-bags: $X_i = \{X_{ij}\}_{j=1}^{N_i}$, and a sub-bag X_{ij} a set of component instances: $X_{ij} = \{x_{ijk}\}_{k=1}^{N_{ij}}$. The Noisy-OR model of MIL in Eq. (3) can be expanded into

(a) (b)

Fig. 3: Examples of candidate positive instances depicted by blue rectangles in (a) and candidate negative instances depicted by red rectangles in (b).

$$p(y_i \mid X_i) = 1 - \prod_j (1 - p(y_{ij} \mid X_{ij})), \tag{4}$$

$$p(y_{ij} \mid X_{ij}) = 1 - \prod_k (1 - p(y_{ijk} \mid x_{ijk})). \tag{5}$$

The likelihood of the positive bag X_i for the component based boosted MIL is expressed by

$$\begin{aligned} p(y_i \mid X_i) &= 1 - \prod_j (1 - (1 - \prod_k (1 - p(y_{ijk} \mid x_{ijk})))) \\ &= 1 - \prod_j (\prod_k (1 - p(y_{ijk} \mid x_{ijk}))) \end{aligned} . \tag{6}$$

The likelihood of each component can be represented by the logistic regression model:

$$p(y_{ij} \mid x_{ij}) = \sigma(H(x_{ij})) = \frac{1}{1 + e^{(-H(x_{ij}))}}. \tag{7}$$

Under the tracking-by-detection paradigm for component based boosted MIL tracker, we need to update the classifier in every frame with selected positive and negative bags. In every frame we detect the position of image patch at which the output of the proposed boosted MIL classifier, learned in the previous frame, is maximized. The optimal detected position $\hat{p}$ of image patch is computed by

$$\hat{p} = \underset{p \in S}{\operatorname{argmax}}\, p(y \mid X^p), \tag{8}$$

where S is a predefined search area, X^p is the image patch at location p, and $p(y \mid X^p)$ is the output of proposed boosted MIL classifier. For collecting learning samples, given the estimated optimal position $\hat{p}$, we collect positive instances (image patches) whose center positions are located in S; negative bags are generated from image patches whose center positions are located outside of S. Figure 3 shows an example of collecting positive instances around the human and negative instances from the background area.

As shown in Algorithm 1, we extend the MILBoost [20] into component based boosted MIL. Given a pool of K candidate weak classifiers h, the algorithm chooses M weak classifiers $\mathbf{h}$ from the candidate pool by optimizing a specific objective function J,

$$\mathbf{h}_k = \underset{h \in \{h_i\}_{i=1}^{K}}{\operatorname{argmax}}\, J(\mathbf{H}_{k-1} + h), \tag{9}$$

Fig. 4: Comparison of tracking results by MILBoost (yellow line) and the proposed component based boosted MIL (blue line). From the first row, ETH Sunny day, UIUC1, skating, and gymnastic athlete sequence are used for comparing results.

Table 1: A quantitative evaluation by comparing mean pixel error. The mean pixel error was computed by measuring the distance between the center position of the bounding box and that of the human being tracked which is manually picked.

Video Sequence		ETH sunny day	UIUC1	Skating	Gymnastics
Mean Pixel Error	MILBoost	13.7	9.57	25.68	35.65
	Component based boosted MIL	9.1	9.46	10.86	19.26

where $\mathbf{H}_{k-1}$ is the strong classifier up to the first $(k-1)$ weak classifiers. The M weak classifiers are selected by noisy-OR model of probability of component labeling rather than those of image patches.

3. EXPERIMENTAL VALIDATION

Each weak classifier h_k is composed of the log odd ratio, $h_k(x) = \log(p(y = 1 \mid x)) - \log(p(y = 0 \mid x))$. We assume that both $p(y = 1 \mid x)$ and $p(y = 0 \mid x)$ are of a normal distribution, $N(\mu_1, \sigma_1)$ and $N(\mu_0, \sigma_0)$, respectively, and their parameters are updated online. We used the turbopixel algorithm [23] for extracting components from each image patch. Harr-like features are extracted in each component. Fig. 4 shows the qualitative results of tracking humans in four video sequences, namely (1) ETH sunny day sequence[1], (2) UIUC standing to sit sequence[2], (3) Skating sequence from VTD webpage[3], and (4) gymnastic athlete sequence from Youtube[4]. We compare our proposed algorithm with MILBoost [20]. Table 1 shows the quantitative results clearly indicating that the proposed algorithm achieves better performance for tracking humans with large articulated motions.

4. CONCLUSION

In this paper, we presented a component based version of the boosted MIL algorithm for tracking articulated human movements. Compared to the MILBoost algorithm, our proposed method can better handle large articulated motions that cause significant variations in the appearance of the humans that need to be tracked. Components in each image patch that are automatically generated by image segmentation provide the additional degrees of freedom that the tracker needs to deal with the large frame-to-frame variations caused by articulated movements. We validated the advantages of our proposed method by comparing the tracking results against the MILBoost algorithm on four different video sequences.

[1] http://www.vision.ee.ethz.ch/~aess/dataset
[2] http://vision.cs.uiuc.edu/projects/activity
[3] http://cs.snu.ac.kr/research/~vtd
[4] http://www.youtube.com/watch?v=FJIrHgshB20

5. REFERENCES

[1] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 34, no. 4, pp. 743–781, 2012.

[2] M. Bertozzi, A. Broggi, A. Fascioli, and A. Tibaldi, "Pedestrian localization and tracking system with kalman filtering," 2004, pp. 584–589.

[3] G. Grubb, A. Zelinsky, L. Nilsson, and M. Rilbe, "3d vision sensing for improved pedestrian safety," in *Intelligent Vehicles Symposium, 2004 IEEE*, june 2004, pp. 19 – 24.

[4] J. Giebel, D. M. Gavrila, and C. Schnorr, "A bayesian framework for multi-cue 3d object tracking," in *In Proceedings of European Conference on Computer Vision*, 2004, pp. 241–252.

[5] K. Wnuk and S. Soatto, "Multiple instance filtering," in *Proc. of NIPS*, 2011, vol. 65, p. 71.

[6] Chad Aeschliman, Johnny Park, and Avinash C. Kak, "A probabilistic framework for joint segmentation and tracking," in *Computer Vision and Pattern Recognition, 2010. CVPR 2010. IEEE Conference on*, 2010.

[7] Yanli Li, Zhong Zhou, and Wei Wu, "Iterative pedestrian segmentation and pose tracking under a probabilistic framework," in *IEEE International Conference on Robotics and Automation*, 2012.

[8] S. Gammeter, A. Ess, T. Jaggli, K. Schindler, B. Leibe, and L. Van Gool, "Articulated multi-body tracking under egomotion," in *European Conference on Computer vision*, 2008.

[9] David Geronimo, Antonio M Lopez, Angle D. Sappa, and Thorsten Graf, "Survey of pedestrian detection for advanced drive assistance systems," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1239–1258, 2010.

[10] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[11] Michael D. Breitenstein, Fabian Reichlin, Bastian Leibe, Esther Koller-Meier, and Luc Van Gool, "Online multiperson tracking-by-detection form a single, uncalibrated camera," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1820–1833, 2011.

[12] Zhe Lin, Gang Hua, and L.S. Davis, "Multiple instance feature for robust part-based object detection," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 405–412.

[13] Li Zhang, Bo Wu, and Ram Nevatia, "Detection and tracking of multiple humans with extensive pose articulation," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.

[14] Bo Yang, Chang Huang, and Ram Nevatia, "Extensive articulated human detection by voting cluster boosted tree," in *Applications of computer vision 2009 Workshop on*. IEEE, 2009, pp. 1–8.

[15] Thomas G. Dietterich, Richard H. Lathrop, and Tomas Lozano-Perez, "Solving the multiple-instance problem with axis-paralle rectangles," *Artificial Intelligence*, vol. 89, pp. 31–71, 1997.

[16] Oded Maron and Tomas Lozano-Perez, "A framework for multiple-instance learning," in *Advances in Neural Information Processing Systems*. 1998, pp. 570–576, MIT Press.

[17] Yixin Chen, Jinbo Bi, and James Z. Wang, "Miles: Multiple-instance learning via embedded instance selection," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 1931–1947, 2006.

[18] Zhouyu Fu, Antonio Robles-Kelly, and Jun Zhou, "Milis: Multiple instancle learning with instance selection," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 958–977, 2011.

[19] Paul Viola, John C. Platt, and Cha Zhang, "Multiple instance boosting for object detection," in *Proc. Neural Information Processing System*, 2005, pp. 1417–1426.

[20] Borid Babenko, Ming-Hsuan Yang, and Serge Belongie, "Robust object tracking with online multiple instance learning," *IEEE Transaction on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1619–1632, 2010.

[21] Helmut Grabner, Michael Grabner, and Horst Bischof, "Real-time tracking via on-line boosting," in *Proc. Conf. British Machine Vision*, 2006, pp. 47–56.

[22] Pramod Sharma, Chang Huang, and Ram Nevatia, "Unsupervised incremental learning for improved object detection in a video," in *International Conference on Computer Vision and Pattern Recognition*, 2012.

[23] Alex Levinshtein, Adrian Stere, Kiriakos N. Kutulakos, David J. Fleet, Sven J. Dickinson, and Kaleem Siddiqi, "Turbopixels: Fast superpixels using geometric flows," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, pp. 2290–2297, 2009.