Structured Learning for Multiple Object Tracking

Wang Yan
wy109@cs.rutgers.edu
Xiaoye Han
xiaoye@cs.rutgers.edu
Vladimir Pavlovic
vladimir@cs.rutgers.edu

Department of Computer Science,
Rutgers, The State University of
New Jersey
USA

Adaptive tracking-by-detection methods use previous tracking results to generate a new training set for object appearance, and update the current model to predict the object location in subsequent frames. Such approaches are typically bootstarpped by manual or semi-automatic initialization in the first several frames. However, most adaptive tracking-by-detection methods focus on tracking of a single object or multiple unrelated objects. Although one can trivially engage several single object trackers to track multiple objects, such solution is frequently suboptimal because it does not utilize the inter-object constraints or the obejct layout information [2].

We propose in this paper an adaptive tracking-by-detection method for multiple objects, inspired by recent work in [1] and [2]. The constraints for structured Support Vector Machine (SVM) in [1] are modified to localize multiple objects simultaneously with both appearance and layout information. Moreover, additional binary constraints are introduced to detect the existences of respective objects and to prevent possible model drift. Thus the method can handle frequent occlusion in multiple object tracking, as well as objects entering or leaving the scene. Those binary constraints make the optimization problem significantly different from the original Structured SVM [3]. The inter-object constraints, embedded in a linear programming technique similar to [2] for optimal position assignment, are applied to diminish false detections.

In single object tracking case, given a set of frames $\{x_1, x_2, \ldots, x_n\}$ indexed by time, and the corresponding set of labeling, i.e. bounding box, $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n\}$, structured SVM tries to find a model $f(x, \mathbf{y})$, such that the task of predicting object location in a testing frame x could be conquered by maximizing:

$$f(x, \mathbf{y}) = \langle \mathbf{w}, \Psi(x|_{\mathbf{y}}) \rangle, \tag{1}$$

where $x|_{\mathbf{y}}$ is the patch of frame x within bounding box $\mathbf{y}$ or the features extract from it, and $\Psi(\cdot)$ is the mapping from input space to the implicit features space. The resulted optimization problem is the following,

$$\min_{\mathbf{w}, \xi} \frac{1}{2}\|\mathbf{w}\|^2 + C_1 \sum_{i=1}^{n} \xi_i \qquad \text{s.t.} \tag{2}$$

$$\langle \mathbf{w}, \Psi(x_i|_{\mathbf{y}_i}) - \Psi(x_i|_{\mathbf{y}}) \rangle \geq \Delta(\mathbf{y}_i, \mathbf{y}) - \xi_i \quad , \quad i = 1, 2, \ldots, n, \quad \mathbf{y} \neq \mathbf{y}_i \tag{3}$$

where $\xi_i \geq 0$, $\mathbf{y} \neq \mathbf{y}_i$ implies bounding box $\mathbf{y}$ in (3) could be anywhere else other than groundtruth $\mathbf{y}$, and $\Delta(\mathbf{y}_i, \mathbf{y}) = 1 - \frac{\mathbf{y}_i \cap \mathbf{y}}{\mathbf{y}_i \cup \mathbf{y}}$ is the loss of predicting $\mathbf{y}$ when groundtruth is $\mathbf{y}_i$.

The tracker could track slowly changing object due to its adaptive nature, but it is also likely to drift when the object is occluded or out of the scene. For selective adaptation and suppression of drifting, binary constraints are added. Suppose Z is the training set of the object detector, and each $\mathbf{z} \in Z$ has the label $l_{\mathbf{z}} \in \{+1, -1\}$. For each $\mathbf{z} \in Z$, the binary constraint is

$$l_{\mathbf{z}}(\langle \mathbf{w}, \Psi(\mathbf{z}) \rangle + b) \geq 1 - \eta_{\mathbf{z}} \quad , \quad \forall \mathbf{z} \in Z, \tag{4}$$

where b is the bias and $\eta_{\mathbf{z}} \geq 0$. (4) favors the sample $\mathbf{z}$ which is correctly classified by current model. The overall objective function (2) becomes

$$\min_{\mathbf{w}, b, \xi, \eta} \frac{1}{2}\|\mathbf{w}\|^2 + C_1 \sum_{i=1}^{n} \xi_i + C_2 \sum_{\mathbf{z} \in Z} \eta_{\mathbf{z}}. \tag{5}$$

Objective function (5), constraints (3)(4) and slack variable constraints lead to a new optimization problem, which could be recognized as a combination of structured SVM and binary SVM.

In multiple object case, compared with the single object version, we add constraint that two or more objects can not appear in the same location in one frame, as well as the objects layout information. The training set includes a frame set $\{x_1, x_2, \ldots, x_n\}$ indexed by time, and $\{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n\}$ is the correspond set of structured labels, where $\mathbf{Y}_i = (\mathbf{y}_i^{(1)}, \mathbf{y}_i^{(2)}, \ldots, \mathbf{y}_i^{(K)})$ indicates the bounding boxes corresponding to K objects in frame i. If the k-th object does not appear in the i-th frame, $\mathbf{y}_i^{(k)} = null$. We design a function $f(x, \mathbf{Y})$ such that the object locations $\mathbf{Y}^*$ in frame x are given by maximizing

$$f(x, \mathbf{Y}) = \sum_{k=1}^{K} \langle \mathbf{w}^{(k)}, \Psi(x|_{\mathbf{y}^{(k)}}) \rangle + \langle \mathbf{v}, \Phi(\mathbf{Y}; \mathbf{Y}_{i-1}) \rangle, \tag{6}$$

where $\mathbf{Y}_{i-1}$ is the layout in $i-1$-th frame and $\Phi(\mathbf{Y}; \mathbf{Y}_{i-1})$ is the layout feature of size $\binom{K}{2} \times 2$, whose k-l-j-th element is

$$\Phi_{klj}(\mathbf{Y}; \mathbf{Y}_{i-1}) = \begin{cases} \left| \left(\mathbf{y}_{i-1}^{(k)}(j) - \mathbf{y}_{i-1}^{(l)}(j) \right) - \left(\mathbf{y}^{(k)}(j) - \mathbf{y}^{(l)}(j) \right) \right| & \text{if } \mathbf{y}_{i-1|i}^{(k|l)} \neq null \\ 0 & \text{otherwise} \end{cases}, \tag{7}$$

while $\mathbf{y}(1)$ and $\mathbf{y}(2)$ are the horizontal and vertical coordinates of the bounding box $\mathbf{y}$'s center, respectively. The model leads the following optimization.

$$\min_{\mathbf{w}, \mathbf{v}, \xi, \eta} \frac{1}{2}\left(\sum_{k=1}^{K} \|\mathbf{w}^{(k)}\|^2 + \|\mathbf{v}\|^2 \right) + C_1 \sum_{i=2}^{n} \xi_i + C_2 \sum_{k=1}^{K} \sum_{\mathbf{z} \in Z} \eta_{\mathbf{z}} \quad \text{s.t.} \tag{8}$$

$$\sum_{k=1}^{K} \langle \mathbf{w}^{(k)}, \Psi(x_i|_{\mathbf{y}_i^{(k)}}) - \Psi(x_i|_{\mathbf{y}^{(k)}}) \rangle + \langle \mathbf{v}, \Phi(\mathbf{Y}_i; \mathbf{Y}_{i-1}) - \Phi(\mathbf{Y}; \mathbf{Y}_{i-1}) \rangle \geq \Delta^M(\mathbf{Y}_i, \mathbf{Y}) - \xi_i, \forall i, \mathbf{Y} \neq \mathbf{Y}_i \tag{9}$$

$$l_{\mathbf{z}^{(k)}}(\langle \mathbf{w}^{(k)}, \Psi(\mathbf{z}^{(k)}) \rangle + b^{(k)}) \geq 1 - \eta_{\mathbf{z}^{(k)}}, \quad \forall k, \quad \forall \mathbf{z}^{(k)} \in Z^{(k)}, \tag{10}$$

(9) is the structured constraint, where $\mathbf{Y}_i$ is the groundtruth object location set of frame i, $\mathbf{Y}$ is the set of locations other than groundtruth, and $\Delta^M(\mathbf{Y}_i, \mathbf{Y}) = \sum_{k=1}^{K} \Delta(\mathbf{y}_i^{(k)}, \mathbf{y}^{(k)})$ is a combination of losses on each objects. (10) is the binary constraint.

Figure 1: Tracking results.

Fig.1 shows the tracking results across time on 2 different video clips, 'motinas-multi-face-fast' and 'toys' respectively. We compared against 5 methods. Evaluation results show that our method works better than other adaptive single object tracker when tracking multiple objects, and outperforms non-adaptive association-based multiple object tracking methods when tracking objects without enough training samples.

[1] S. Hare, A. Saffari, and P.H.S. Torr. Struck: Structured output tracking with kernels. In *IEEE International Conference on Computer Vision*, 2011.

[2] H. Jiang, F. Fels, and J. Little. A linear programming approach for multiple object tracking. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2007.

[3] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *International Conference on Machine Learning*, 2004.