Online Motion Agreement Tracking

Zheng Wu
wuzheng@bu.edu
Jianming Zhang
jmzhang@bu.edu
Margrit Betke
betke@bu.edu

Computer Science Department
Boston University
Boston, USA

We present a new appearance model for online multi-target tracking, where a patch-based representation of each target to be tracked is introduced along with a sequential update scheme, which we call "motion agreement tracking" (MAT). This method updates the appearance model of each target online by indirectly evaluating the motion consistency among its local patches. We show that a distance measure based on appropriately re-weighted local patches will successfully reduce tracker errors especially those lead to track fragmentation and track switching. We also demonstrate that the performance results of a well-designed online 2D tracker, like the MAT algorithm, can actually measure up to those of state-of-the-art offline algorithms on various popular pedestrian tracking benchmarks. Our competitive results are particularly appealing since our technique is very efficient. They also suggest that the role of a proper appearance model may be more important than researchers deemed for video-based tracking, where the majority of previous studies focused on motion dynamics.

In essence, we designed a person-specific appearance model with a collection of local image patches by dividing the bounding box of a detected person into a grid representation. Each local patch k is described by a 64-bin color histogram in HSV space. Each patch is associated with a weight w_k, which is set to be uniform when the tracker is initialized. We would like to assign a high weight to a stable patch that does not change significantly over time, and a low weight that the patch belongs to the background or represents a fast-changing part of the object.

When a detection is assigned to the tracker after solving the tracker-measurement assignment problem, a filtered estimate of the object's global motion vector v is computed by a Kalman filter. At the same time, the method estimates the local displacement v_k of each patch based on a similarity measure. Here, we compare two popular measures in our system: the maximum normalized cross-correlation and the minimum histogram intersection distance.

Given the local motion estimates, our method re-weights each patch by checking the agreement between v_k and the global motion v. The intuition is that if v_k is similar to v, then this local patch moves along with the pedestrian, so it is more likely to be a stable region that does not undergo significant appearance change. By focusing our effort on the most stable patches, we can construct a similarity measure that can distinguish between interacting objects. The discrete level g of the agreement is computed by our implementation as follows:

$$\begin{aligned} \theta_k &= \cos^{-1}\left(\frac{v_k \cdot v}{\|v_k\|\|v\|}\right) \\ g_k &= \begin{cases} 2, & \text{if } \theta_k < \frac{\pi}{4} \\ 1, & \text{if patch is in an occlusion relationship} \\ 0, & \text{otherwise} \end{cases} \end{aligned} \quad (1)$$

where θ_k is the angle between the two velocity vectors v and v_k. The motion score g is defined to be symmetric on both agreement and disagreement sides; the magnitude is not important here, as its contribution to the following weight update will be normalized. A score of $g_k = 1$ suggests a random guess. Using g_k, our algorithm updates the weight w_k associated with each patch k at time t as follows:

$$\begin{aligned} w_k(t) &= \frac{\alpha^{(t-1)} w_k^{(t-1)} + g_k,}{\alpha^{(t-1)} + \sum_k g_k} \\ \alpha^{(t)} &= \frac{\alpha^{(t-1)} + \sum_k g_k}{2}, \end{aligned} \quad (2)$$

where α is a self-adaptive learning rate which controls how much the current estimates influence the update. A large value of α suggests a smooth update of weights at the current frame, which happens when most local motion estimates agree with the global motion model. Both α and w are non-negative, and w is always normalized.

Figure 1: Sample images from our tracking results. Colors and numbers indicate tracks corresponding to different people.

Finally, the distance function between an object i and a candidate detection j in the data association step is given as:

$$c_{i,j} = (1 - \frac{|b_i \bigcap b_j|}{|b_i \bigcup b_j|}) + \lambda \sum_k^{\#\text{patches}} w_k (1 - \sum_n^{\#\text{bins}} \min(h_i^n, h_j^n)), \quad (3)$$

where the first term evaluates the alignment of two bounding boxes i and j, and the second term measures the re-weighted difference of the appearance of i and j based on the histogram intersection distance. Parameter λ balances the contributions of the motion and appearance terms.

We integrated our appearance model into an online tracking system and tested it extensively on nine tracking benchmarks. The performance was evaluated by comparing our MAT method with a simple baseline, an online tracking algorithm that models the appearance holistically [5] and four offline tracking algorithms [1, 2, 3, 4] that solve the data association globally. Our results suggest that our online MAT algorithm indeed outperforms state-of-the-art algorithms for various benchmark videos. Given its efficiency and ease-of-use, our MAT algorithm is even valuable for tracking scenarios where its performance is expected to be inferior because the tracks it produces online may be used as valuable initializations for offline tracking algorithms.

[1] A. Andriyenko, K. Schindler, and S. Roth. Discrete-continuous optimization for multi-target tracking. In *CVPR*, 2012.

[2] C.-H. Kuo and R. Nevatia. How does person identity recognition help multi-person tracking. In *CVPR*, 2011.

[3] H. Pirsiavash, D. Ramanan, and C. Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *CVPR*, 2011.

[4] B. Yang and R. Nevatia. An online learned CRF model for multi-target tracking. In *CVPR*, 2012.

[5] J. Zhang, L. Lo Presti, and S. Sclaroff. Online multi-person tracking by tracker hierarchy. In *Proceeding of the IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS)*, 2012.