Multiple Object Tracking Using Local Motion Patterns

Mehrsan Javan Roshtkhari
http://www.cim.mcgill.ca/~javan
Martin D. Levine
http://www.cim.mcgill.ca/~levine

Center For Intelligent Machines
Department of Electrical and Computer Engineering
McGill University
Montreal, QC, Canada

Object tracking is, perhaps, the most fundamental task for any high-level video content analysis system. Decades of research on this topic have produced a diverse set of approaches and a rich collection of tracking algorithms. Most of the reported algorithms are based on object detection followed by a data association algorithm. Thus a key assumption is that a reliable object detection algorithm exists [1, 5]. These methods use the detection response to construct an object trajectory. This is accomplished by using data association based on either the detection responses or a set of short tracks called tracklets that are associated with each detected object [1]. Subsequently, data association links these tracklets into multi-frame trajectories. On the other hand, there are other tracking algorithms, which are based on local spatio-temporal motion patterns in the scene. More closely related to our approach are those that construct motion models for the moving objects without performing any detection [2].

In this paper we concentrate on creating long-term trajectories for unknown moving objects by using a model-free tracking algorithm. As opposed to the tracking-by-detection algorithms [5], no object detection is involved. Each individual object is tracked only by modeling the temporal relationship between sequentially occurring local motion patterns. This is achieved by constructing two sets of initial tracks that code local and global motion patterns in videos. These local motion patterns are obtained by analyzing spatially and temporally varying structures in videos [3, 4].

Initially, the video is densely sampled, spatio-temporal video volumes (STVs) are constructed, and similar ones are grouped to reduce the dimension of the search space. This is called the low-level codebook, $\mathbf{C}^{\mathcal{L}}$. Then, a large contextual region containing many STVs (in space and time) around each pixel is examined and their compositional relationships are approximated using a probabilistic framework. They are then employed to form yet another codebook, called the high-level codebook, $\mathbf{C}^{\mathcal{H}}$. Therefore, two codewords are assigned to each pixel, one from the low level and the other from the high level codebook. By examining pairs of sequential video frames, the matching codewords for each video pixel are transitively linked into distinct tracks, whose total number is unknown a priori and which we will refer to as linklets. The linking process is separately performed for both codebooks. This is done under the hard constraint that no two linklets may share the same pixel at the same time, i.e. the assigned codewords. The end result at this step is two sets of independent linklets obtained from the low- and high-level codebooks.

Subsequently, a set of sparse tracks, referred to as tracklets in the literature, are produced by grouping the linklets that indicate similar motion patterns (see Figure 1). This produces two sets of independent tracklets, referred to as low- and high-level tracklets, $\mathbf{T}^{\mathcal{L}}$ and $\mathbf{T}^{\mathcal{H}}$, respectively. Given the resulting tracklets, high-level trajectories can be generated by linking them in space and time. We achieve this by formulating the data association required as a maximum a posteriori (MAP) problem and solve it with the Markov Chain Monte Carlo Data Association (MCMCDA) algorithm. The observations are taken to be the constructed tracklets, $\mathcal{O} = \left\{\mathbf{T}^{\mathcal{L}}, \mathbf{T}^{\mathcal{H}}\right\}$. Let $\mathbf{\Gamma}$ be a tracklet association result, which is a set of trajectories, $\Gamma_k \in \mathbf{\Gamma}$. Γ_k is defined as a set of the connected observations which is a subset of all observations, $\Gamma_k = \left\{T_i^{\mathcal{L}}, T_j^{\mathcal{H}}\right\} \subseteq \mathcal{O}$. The goal is to find the most probable set of object trajectories, $\mathbf{\Gamma}$, which is formulated as a MAP problem:

$$\mathbf{\Gamma}^* = \arg\max_{\mathbf{\Gamma}} P(\mathbf{\Gamma}|\mathcal{O}) = \arg\max_{\mathbf{\Gamma}} P(\mathcal{O}|\mathbf{\Gamma})\,P(\mathbf{\Gamma}) \tag{1}$$

The likelihood, $P(\mathcal{O}|\mathbf{\Gamma})$ indicates how well a set of trajectories matches the observations and the prior, $P(\mathbf{\Gamma})$ indicates how correct the data association is. By assuming that the likelihoods of the tracklets are conditionally independent, we can rewrite the likelihood, $P(\mathcal{O}|\mathbf{\Gamma})$, in (1) as follows:

$$P(\mathcal{O}|\mathbf{\Gamma}) = \prod_{\substack{T_i^{\mathcal{L}} \in \mathbf{T}^{\mathcal{L}} \\ T_j^{\mathcal{H}} \in \mathbf{T}^{\mathcal{H}}}} P\left(T_i^{\mathcal{L}}, T_j^{\mathcal{H}}|\mathbf{\Gamma}\right) \prod_{\Gamma_k \in \mathbf{\Gamma}} P(\Gamma_k) \tag{2}$$

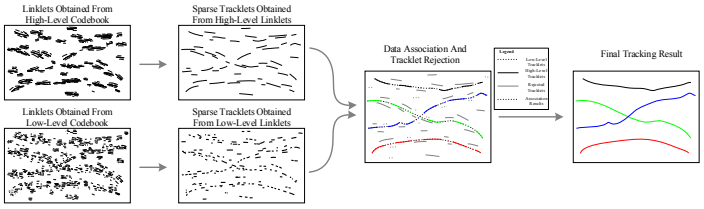

Figure 1: The goal is to estimate the trajectory of the moving objects in the video without invoking object detection. Initially two sets of linklets are constructed by chaining; the low-level considers small window fragments, while the high-level analyzes a larger region in order to impose a contextual influence. They are obtained by exploiting an activity understanding system [3, 4]. The resultant tracks (chains) are filtered and replaced by a set of sparse representative tracks, the so-called tracklets. Longer trajectories are then generated by using the Markov Chain Monte Carlo Data Association (MCMCDA) algorithm to solve the Maximum A Posteriori (MAP) problem using tracklet affinities. Thus this procedure uses low-level tracklets to connect high-level tracklets when there is a discontinuity in motion or time.

We adopt Markov Chain Monte Carlo Data Association (MCMCDA) to estimate an initially unspecified number of trajectories. To this end, we formulate the tracklet association problem as a Maximum A Posteriori (MAP) problem to produce a chain of tracklets. Data association is accomplished by considering temporal continuity and motion consistency of both the low- and high-level tracklets, with the additional option of rejecting irrelevant tracklets. The final output of the data association algorithm is a partition of the set of tracklets such that those belonging to each individual object have been grouped together. Implementation of this method is described in the paper, as are the details of the all other parts of this algorithm.

Although our algorithm possesses no information regarding either an object's color pattern or a human body model, it achieves promising results on challenging data sets. The results indicate that although the correct detections we obtain with our algorithm are comparable to the state of the art, they include more false positives. Perhaps one can expect this, since no object detection is employed in our algorithm. Recall that the scene observations that we use are motion descriptors and do not incorporate object appearance, as do object-centric trackers. As stated in the paper, the major drawback of our algorithm is the number of false positives and some problems in maintaining the trajectory identity when objects have similar shape and motion.

[1] Huang Chang, Li Yuan, and R. Nevatia. Multiple target tracking by learning-based hierarchical association of detection responses. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35 (4):898–910, 2013.

[2] L. Kratz and K. Nishino. Tracking pedestrians using local spatio-temporal motion patterns in extremely crowded scenes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(5):987–1002, 2012.

[3] Mehrsan Javan Roshtkhari and Martin D. Levine. Online dominant and anomalous behavior detection in videos. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2609–2616, 2013.

[4] Mehrsan Javan Roshtkhari and Martin D. Levine. Human activity recognition in videos using a single example. *Image and Vision Computing*, 31(11):864–876, 2013.

[5] Bo Yang and Ramakant Nevatia. Multi-target tracking by online learning a crf model of appearance and motion patterns. *International Journal of Computer Vision*, 107(2):203–217, 2014.