A Detection-Based Multiple Object Tracking Method

Mei Han Amit Sethi[†] Yihong Gong
NEC Laboratories America, Cupertino, CA, USA
[†]University of Illinois at Urbana Champaign, Champaign, IL, USA

Abstract

In this paper we describe a method for tracking multiple objects whose number is unknown and varies during tracking. Based on preliminary results of object detection in each image which may have missing and/or false detection, the multiple object tracking method keeps a graph structure where it maintains multiple hypotheses about the number and the trajectories of the objects in the video. The image information drives the process of extending and pruning the graph, and determines the best hypothesis to explain the video. While the image-based object detection makes a local decision, the tracking process confirms and validates the detection through time, therefore, it can be regarded as temporal detection which makes a global decision across time. The multiple object tracking method gives feedbacks which are predictions of object locations to the object detection module. Therefore, the method integrates object detection and tracking tightly. The most possible hypothesis provides the multiple object tracking result. The experimental results are presented.

1 Introduction

Multiple object tracking has been a challenging research topic in computer vision. It has to deal with the difficulties existing in single object tracking, such as changing appearances, non-rigid motion, dynamic illumination and occlusion, as well as the problems related to multiple object tracking including inter-object occlusion, multi-object confusion. There has been much work on multiple object visual tracking. MacCormick and Blake [1] use a sampling algorithm for tracking fixed number of objects. Tao et al. [2] present an efficient hierarchical algorithm to track multiple people. Isard and MacCormick [3] propose a Bayesian multiple-blob tracker. Hue et al. [4] describe an extension of classical particle filter where the stochastic assignment vector is estimated by a Gibbs sampler. These methods only keep one hypothesis of the tracking result which has the largest posterior probability based on current and previous observations. They may fail with background clutter, occlusion and multi-object confusion. Multiple hypothesis methods are more robust because the tracking result corresponds to the state sequence which maximizes the joint state-observation probability.

A well-known early work in multiple hypothesis tracking (MHT) is the algorithm developed by Reid [5]. The joint probabilistic data association filter (JPDAF) [6] finds the state estimate by evaluating the measurement-to-track association probabilities. Some methods [7, 8] are presented to model the data association as random variables which are estimated jointly with state estimation by EM iterations. Most of these work are in the small target tracking community where object representation is simple.

We propose a multiple hypothesis method to track multiple objects based on object detection. The detection results recognize the tracking targets in each image. Any object detection method can be used. In our implementation, we apply a neural network based object detection module to detect pedestrians. The tracking algorithm accumulates the detection results in a graph-like structure and maintains multiple hypotheses of objects trajectories. At the same time, the multiple object tracking method gives feedbacks which are predictions of object locations to the object detection module. Therefore, the tracking method tightly integrates object detection and tracking to guarantee a robust and efficient tracking algorithm. Many people have worked on the integration of object detection and tracking. SVM tracker [9] applies recognition algorithms to efficient visual tracking. Many systems of multiple people detection and tracking are presented using aspect ratio [10], silhouette [11], human shape model [12] to detect human. None of these methods maintains multiple hypotheses.

Our multiple object tracking method is reliable to deal with occlusions, irregular object motions, changing appearances by postponing the decision of object trajectories until sufficient information is accumulated over time. It makes a global decision. The most possible hypothesis generates the multiple object tracking result. The trajectories provide information of object identifications, motion histories, timing and object interactions. The information can be applied to detect abnormal behaviors in video surveillance and collect traffic data in traffic control systems.

2 Object Detection

The multiple object tracking method works on fixed cameras. It starts with an adaptive background modelling module which deals with changing illuminations and does not require objects to be constantly moving. A Gaussian-mixture

(a) (b) (c)

Figure 1: Object detection: (a) original image, (b) foreground mask image, the white pixels represent the mask of the foreground objects, (c) human detection results, the lighter pixels show the higher detection probabilities.

based background modelling method [13] is used to generate a binary foreground mask image as shown in Figure 1(b). The white pixels represent the mask of the foreground objects. An object detection module takes the foreground pixels generated by background modelling as input and outputs the probabilities of object detection. It searches over the foreground pixels and gives the probability of each location where a certain scale object is found. Any object detection approach can be fit into this part. In our implementation, we apply a neural network based object detection module to detect pedestrians. Each foreground blob is potentially the image of a person. Each pixel location is applied to a neural network that has been trained for this task. The neural network generates a score, or probability, indicative of the probability that the blob around the pixel does in fact represent a human of some scale. A particular part of the detected person, e.g., the approximate center of the top of the head, is illustratively used as the "location" of the object, which is shown as a light spot in Figure 1(c). The lighter spot demonstrates the higher detection score. The neural network searches over each pixel at a few scales. The detection score corresponds to the best score, i.e., the largest detection probability, among all scales.

3 Tracking Algorithm

The tracking algorithm accepts the probabilities of preliminary object detection and keeps multiple hypotheses of object trajectories in a graph structure, as shown in Figure 2. Each hypothesis consists of the number of objects and their trajectories. The first step in tracking is to extend the graph to include the most recent object detection results, that is, to generate multiple hypotheses about the trajectories. An image based likelihood is then computed to give a probability to each hypothesis. This computation is based on the object detection probability, appearance similarity, trajectory smoothness and image foreground coverage and compactness. The probabilities are calculated based on a sequence of images, therefore, they are temporally global representations of hypotheses likelihood. The hypotheses are ranked by their probabilities and the unlikely hypotheses are pruned from the graph in the hypotheses-management step. In this way a limited number of hypotheses are maintained in the graph structure, which improves the computation efficiency.

In the graph structure (Figure 2), the graph nodes represent the object detection results. Each node is composed of the object detection probability, object size or scale, location and appearance. Each link in the graph is computed based on position closeness, size similarity and appearance similarity between two nodes (detected objects). The graph is extended over time. In this section we describe three steps of the tracking algorithm: hypotheses generation, likelihood computation and hypotheses management.

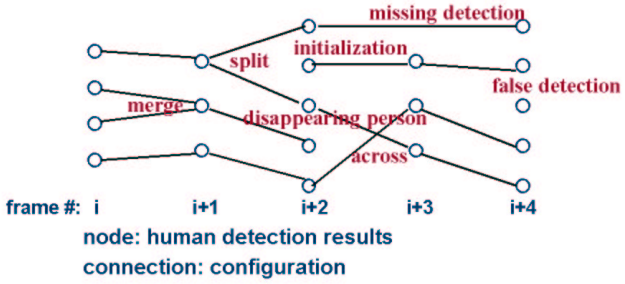


Figure 2: Graph structure in multiple object tracking

3.1 Hypotheses Generation

Given object detection results in each image, the hypotheses generation step firstly calculates the connections between the maintained graph nodes and the new nodes from current image. The maintained nodes include the ending nodes of all the trajectories in maintained hypotheses. They are not necessarily from the previous image since object detection may have missing detections. The connection probability is computed according to,

$$p_{\text{con}} = w_{\text{a}} \times p_{\text{a}} + w_{\text{p}} \times p_{\text{p}} + w_{\text{s}} \times p_{\text{s}} \quad (1)$$

where w_{a}, w_{p} and w_{s} are the weights in the connection probability computation, that is, the connection probability is a weighted combination of appearance similarity probability p_{a}, position closeness probability p_{p} and size similarity probability p_{s}. We prune the connections whose probabilities are very low for the sake of computation efficiency.

As shown in Figure 2, the generation process takes care of object occlusion by track splitting and merging. When a person appears from occlusion, the occluding track splits into two tracks, on the other hand, when a person gets occluded, the corresponding node is connected (merged) with the occluding node. The generation process deals with missing data naturally by skipping nodes in graph extensions, that is, the connection is not necessarily built on two nodes from consecutive image frames. The generation handles false detections by keeping the hypotheses ignoring some nodes. It initializes new trajectories for some nodes depending on their (weak) connections with existing nodes

and their locations (at appearing areas, such as doors, view boundaries). The multiple object tracking algorithm keeps all possible hypotheses in the graph structure. At each local step, it extends and prunes the graph in a balanced way to maintain the hypotheses as diversified as possible and delays the decision of most likely hypothesis to a later step.

3.2 Likelihood Computation

The likelihood or probability of each hypothesis generated in the first step is computed according to the connection probability, the object detection probability, trajectory analysis and the image likelihood computation. The hypothesis likelihood is accumulated over image sequences,

$$\begin{aligned} \text{likelihood}_i &= \text{likelihood}_{i-1} \\ &+ \frac{\sum_{j=1}^{n} log(p_{\text{con}_j}) + log(p_{\text{obj}_j}) + log(p_{\text{trj}_j})}{n} \\ &+ L_{\text{img}} \end{aligned} \quad (2)$$

where i is the current image frame number, n represents the number of objects in current hypothesis. p_{con_j} denotes the connection probability of jth trajectory computed in Equation (1). If jth trajectory has missing detection in current frame, a small probability, i.e., missing probability, is assigned to p_{con_j}. p_{obj_j} is the object detection probability and p_{trj_j} measures the smoothness of jth trajectory. We use the average likelihood of multiple trajectories in the computation. The metric prefers the hypotheses with better human detections, stronger similarity measurements and smoother tracks. L_{img} is the image likelihood of the hypothesis. It is composed of two items,

$$L_{\text{img}} = l_{\text{cov}} + l_{\text{comp}} \quad (3)$$

where

$$\begin{aligned} l_{\text{cov}} &= log\left(\frac{|A \bigcap (\bigcup_{j=1}^{n} B_j) + c|}{|A| + c}\right) \\ l_{\text{comp}} &= log\left(\frac{|A \bigcap (\bigcup_{j=1}^{n} B_j) + c|}{|\sum_{j=1}^{n} B_j| + c}\right) \end{aligned} \quad (4)$$

l_{cov} calculates the hypothesis coverage of the foreground pixels and l_{comp} measures the hypothesis compactness. A denotes the sum of foreground pixels and B_j represents the pixels covered by jth node (or track). $\bigcap$ denotes the set intersection and $\bigcup$ the set union. The numerators in both l_{cov} and l_{comp} represent the foreground pixels covered by the combination of multiple trajectories in current hypothesis, therefore, l_{cov} represents the foreground coverage of the hypothesis, the higher the larger coverage, and l_{comp} measures how much the nodes overlap with each other, the larger the less overlap and the more compact. c is a constant. These two values give a spatially global explanation of the image (foreground) information. This computation is similar to the image likelihood computation in [2].

The hypothesis likelihood is a value refined over time. It provides a global description of object detection results. Generally speaking, the hypotheses with higher likelihood are composed of better object detections with good image explanation. It tolerates missing and false detections since it has a global view of image sequences.

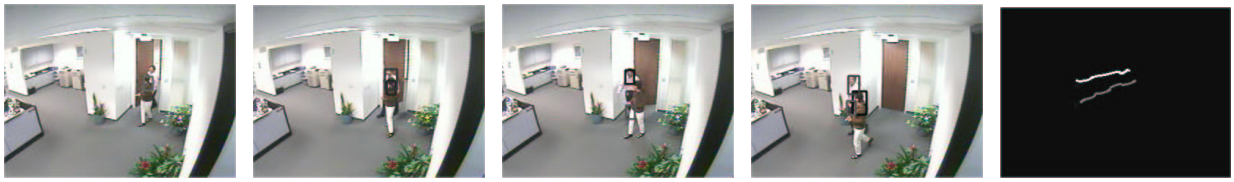
3.3 Hypotheses Management

This step ranks the hypotheses according to their likelihood values. To avoid combinatorial explosion in graph extension, we only keep a limited number of hypotheses and prune the graph accordingly. The hypotheses management step deletes the out-of-date tracks, which correspond to the objects which are gone for a while, and keeps a short list of active nodes which are the ending nodes of the trajectories of all the kept hypotheses. The number of active nodes is the key to determine the scale of graph extension, therefore, a careful management step assures efficient computation. The design of this multiple object tracking algorithm follows two principles: 1. We keep as many hypotheses as possible and make them as diversified as possible to cover all the possible explanations of image sequences. The top hypothesis is chosen at a later time to guarantee it is an informed and global decision. 2. We make local prunes of unlikely connections and keep only a limited number of hypotheses. With reasonable assumptions of these thresholds, the method achieves real-time performance in a not-too-crowded environment. The graph structure is applied to keep multiple hypotheses and make reasonable prunes for both reliable performance and efficient computation.

The tracking module provides feedbacks to the object detection module to improve the local detection performance. According to the trajectories in the top hypothesis, the multiple object tracking module predicts the most likely locations to detect objects. This interaction tightly integrates the object detection and tracking, and makes both of them more reliable.

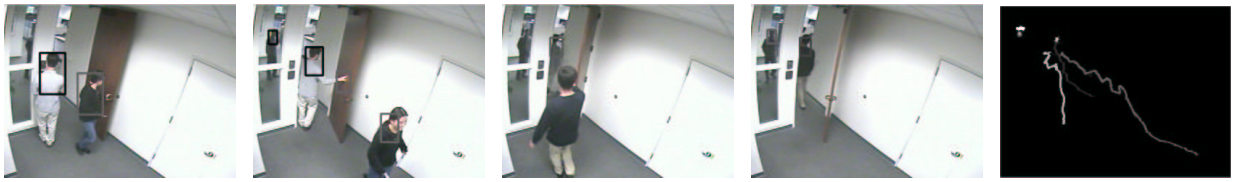
4 Experiment

The multiple object tracking method has been tested on two existing CCTV cameras. The first scenario includes two persons coming into the door about the same time. Figure 3(a) shows 4 images from the sequence with overlaid bounding boxes showing the human detection results. The darker the bound box the higher the detection probability. Figure 3(b) demonstrates the multi-tracks with the largest probability generated by the multiple object tracking. The tracks are overlaid on the detection score map. Different intensities represent different tracks. The human detection based on each image is certainly not perfect. In the first and third images, the human detector misses the person in the back due to occlusion and the person in the front



(a) (b)

Figure 3: Tracking results with missing/false human detections: (a) original images with overlaid bounding boxes showing the human detection results, (b) multiple object tracking result overlaid on the human detection map.



(a) (b)

Figure 4: Tracking results of crossing tracks: (a) original images with overlaid bounding boxes showing the human detection results, (b) multiple object tracking result overlaid on the human detection map.

due to distortion, respectively. There are false detections in the forth image caused by background noise and people interaction. However, the multiple object tracking method manages to maintain the right number of tracks and their configurations, as shown in Figure 3(b), because it searches for the best explanation sequence of the observations over time.

Figure 4 demonstrates an example of multiple people tracking with crossing tracks. The example first shows the lady opens the door for the person in gray shirt, then the person in dark shirt follows and goes into the area. Figure 4(a) shows the images from the sequence and (b) demonstrates the tracking result. Interestingly, there is one short track close to the up-left corner of the result image because one person is standing inside the door and the human detection consistently detects him through the glass window. Therefore, 4 tracks are shown in Figure 4(b), the short track for the standing person, the long track for the lady, the light track for the guy in gray shirt, and the dark track for the guy in dark shirt.

References

[1] J.P. MacCormick and A. Blake, "A probabilistic exclusion principle for tracking multiple objects," in *ICCV99*, 1999, pp. 572–578.

[2] H. Tao, H.S. Sawhney, and R. Kumar, "A sampling algorithm for tracking multiple objects," in *Vision Algorithms 99*, 1999.

[3] M. Isard and J.P. MacCormick, "Bramble: A bayesian multiple-blob tracker," in *ICCV01*, 2001, pp. II: 34–41.

[4] C. Hue, J.P. Le Cadre, and P. Perez, "Tracking multiple objects with particle filtering," *IEEE Trans. on Aerospace and Electronic Systems*, vol. 38, no. 3, pp. 791–812, July 2002.

[5] D.B. Reid, "An algorithm for tracking multiple targets," *AC*, vol. 24, no. 6, pp. 843–854, December 1979.

[6] T.E. Fortmann, Y. Bar-Shalom, and M. Scheffe, "Sonar tracking of multiple targets using joint probabilistic data association," *IEEE Journal Oceanic Eng.*, vol. OE-8, pp. 173–184, July 1983.

[7] R.L. Streit and T.E. Luginbuhl, "Maximum likelihood method for probabilistic multi-hypothesis tracking," in *Proceedings of SPIE International Symposium, Signal and Data Processing of Small Targets*, 1994.

[8] H. Gauvrit and J.P. Le Cadre, "A formulation of multitarget tracking as an incomplete data problem," *IEEE Trans. on Aerospace and Electronic Systems*, vol. 33, no. 4, pp. 1242–1257, Oct 1997.

[9] S. Avidan, "Support vector tracking," in *CVPR01*, 2001, pp. I:184–191.

[10] I. Haritaoglu, D. Harwood, and L.S. Davis, "W4s: A real-time system for detecting and tracking people in 2 1/2-d," in *ECCV98*, 1998.

[11] I. Haritaoglu, D. Harwood, and L.S. Davis, "Hydra: Multiple people detection and tracking using silhouettes," in *VS99*, 1999.

[12] T. Zhao, R. Nevatia, and F. Lv, "Segmentation and tracking of multiple humans in complex situations," in *CVPR01*, 2001, pp. II:194–201.

[13] C. Stauffer and W.E.L. Grimson, "Learning patterns of activity using real-time tracking," *PAMI*, vol. 22, no. 8, pp. 747–757, August 2000.