Detection and Recognition Technologies

Integrated Object Detection and Tracking by Multiple Hypothesis Analysis

By Yihong GONG*

ABSTRACT In this paper, we describe a novel multi-object tracking technique that integrates object detection into the object tracking process and solves the tracking problem by finding the globally optimized object trajectories through the multiple hypothesis analysis. The detection module recognizes the target objects in each frame of the video stream. The tracking module accumulates the detection results in a graph-like structure and maintains multiple hypotheses of objects trajectories. The hypotheses are ranked by their likelihoods which are computed over a sufficient number of frames, and the most likely hypothesis is used to generate the object tracking result. At the same time, the tracking module gives feedbacks to the object detection module, which are predictions of object locations in subsequent frames. Through such tight integration of the object detection and tracking, as well as the global optimization of object trajectories, we have accomplished not only robust and efficient object tracking, but also the ability to deal with occlusions, irregular object motions, changing appearances, etc. which are the challenging problems for most traditional tracking methods.

KEYWORDS **SmartCatch, Intelligent video surveillance, Multiple object tracking, Multiple hypothesis analysis**

1. INTRODUCTION

"9·11" terrorist attacks on the United States have dramatically changed the landscape of the security industry in the world. Nowadays in most countries, millions, or even hundreds of millions of video surveillance cameras have been installed in airports, train stations, shopping malls, government or public facilities, etc., and the enormous volume of video streams generated by these surveillance cameras are inundating security guards in central monitoring facilities. Because of budget restraints, many organizations cannot afford to employ sufficient manpower to actively monitor all the surveillance cameras, and just simply record and archive the video streams onto hard-disks or video tapes instead. As a consequence, these large-scale, expensive surveillance systems are degraded into less valuable video capturing and archiving tools from the effective weapons for discovering crimes on the spot or preventing crimes from happening.

To relieve security guards from the tedious video monitoring task, and to make surveillance systems cost-effective crime-fighting weapons, it becomes urgent and critical to develop intelligent video surveillance technologies that are able to monitor incoming video streams on behalf of human security guards, and to generate alarms whenever security threatening events are detected. In the past two years, at NEC Laboratories America, we developed the SmartCatch video surveillance system for several major airports in the United States. This system can automatically detect a variety of suspicious behaviors and rule violations that may constitute security threats to an airport. The SmartCatch system has been up and running uninterruptedly for more than a year in several customers' sites, and has exceeded the customers' expectations in event detection accuracies.

The distinguishing feature of the SmartCatch system is its ability to accurately detect objects of interest, and to robustly track all the detected objects over long time periods. This ability enables us to detect the appearances and disappearances of objects of interest in the surveillance zone, and to obtain such information for each object as its trajectory, motion history, timing, and interactions with other objects. Based on these data, we can certainly recognize various behaviors, so as to detect the events of potential security threats. The accuracy of object detections and the robustness of multiple object tracking arise from our

*NEC Laboratories America, Inc.

novel technique that integrates object detection into the object tracking process and solves the tracking problem by finding the globally optimized object trajectories through the multiple hypothesis analysis.

In this paper, we present the core technology of the SmartCatch video surveillance system, which is the integration of objection detection and tracking under the framework of multiple hypothesis analysis. In the following part, Section 2 briefly describes related works; Section 3 provides the overview of the core technology; Sections 4 and 5 describe the two main components of the operation: Object detection and tracking; and Section 6 presents experimental results.

2. RELATED WORK

Traditional object tracking methods generally treat object detection as a separate task. The most common approach is that either a human operator or a separate program conducts object detection from the incoming video stream, and initiates the tracking program whenever the target object is detected. Once the target object is passed over to the tracking program, the interaction between the object detector and the tracker is completed, and the object tracking becomes the sole responsibility of the tracking program. A major problem of this approach is that tracking errors may occur at certain frames due to changing appearances, non-rigid motions, and dynamic illuminations of the target object. If there is no self-correction mechanism, these tracking errors will accumulate gradually, and will eventually cause the tracker to drift away from the target object into tracking an irrelevant one. Furthermore, when multiple objects are present in the scene, tracking them simultaneously brings about additional challenges such as inter-object occlusions, multi-object intersections, etc.

There have been some research studies in the literature that attempt to address the above tracking problems. MacCormick and Blake[1] used a sampling algorithm for tracking fixed number of objects. Tao et al.[2] presented an efficient hierarchical algorithm to track multiple people. Isard and MacCormick[3] proposed a Bayesian multiple-blob tracker. Hue et al.[4] described an extension of classical particle filter where the stochastic assignment vector is estimated by a Gibbs sampler. These methods, however, keep only one hypothesis of the tracking result which has the largest posterior probability based on current and previous observations. They may fail with background clutter, occlusions and multi-object confusions. 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 works are in the small target tracking community where object representation is simple.

3. METHOD OUTLINE

We propose a novel multiple object tracking technique that integrates object detection into the object tracking process and discovers the globally optimized object trajectories through the multiple hypothesis analysis. The detection module recognizes the target objects in each frame. Any object detection method can be used here. In our implementation, we applied a neural network-based object detection module to detect pedestrians. The tracking module accumulates the detection results in a graph-like structure and maintains multiple hypotheses of objects trajectories. At the same time, the tracking module gives feedbacks to the object detection module, which are predictions of object locations in subsequent frames. These feedbacks cause the detection module to search over the predicted areas with a higher priority, and to give higher probability scores to the detected objects that match the prediction results.

The object detection result obtained from each frame is used by the tracking module to create new hypotheses, and update the existing hypotheses of object trajectories. Our multiple object tracking method makes a global decision on object trajectories by selecting the most probable hypothesis that is accumulated over a sufficient number of frames. Through such tight integration of the object detection and tracking, as well as the global optimization of object trajectories, we have accomplished not only robust and efficient object tracking, but also the ability to deal with occlusions, irregular object motions, changing appearances, etc. which are the challenging problems for most traditional tracking methods.

The object trajectories provide information of object identifications, motion histories, timing and object interactions. Such information can be applied to detect abnormal behaviors in video surveillance and collect traffic data in traffic control systems.

4. OBJECT DETECTION

The multiple object tracking method starts with an adaptive background modeling module which deals with changing illuminations. A Gaussian mixture-based background modeling method[13] is used to generate a binary foreground bitmap image as shown in **Fig. 1**(b).The white pixels represent the bitmap of the foreground objects. An object detection module takes the foreground bitmap as the input and outputs the probabilities of object detection. Any object detection method can be fit into this part. In our implementation, we applied a neural network-based object detection module to detect pedestrians. The neural network searches over the foreground bitmap and gives each location of the bitmap a probability that a target object is found at this location. As a result, a probability map of the size same as the foreground bitmap is generated by the object detection module where the value at each point of the map indicates the probability that the point contains a target object in the original frame. A particular part of the detected person, e.g., the center of the human head, is illustratively used as the "location" of the object, which is shown as a light spot in Fig. 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.

5. OBJECT TRACKING

The tracking module accepts the probabilities of the preliminary object detection and keeps multiple hypotheses of object trajectories in a graph structure, as shown in **Fig. 2**. Each hypothesis consists of a fixed 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 (Fig. 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 the position closeness, size similarity and appearance similarity between the two nodes (detected objects). The graph is extended over time. In this section we describe the three steps of the tracking module: hypotheses generation, likelihood computation and hypotheses management.

5.1 Hypotheses Generation

Given object detection results in each image, the hypotheses generation step first calculates the connections between the maintained graph nodes and the new nodes from the 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:

$$p_{con} = w_a \cdot p_a + w_p \cdot p_p + w_s \cdot p_s \quad (1)$$

Fig. 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.

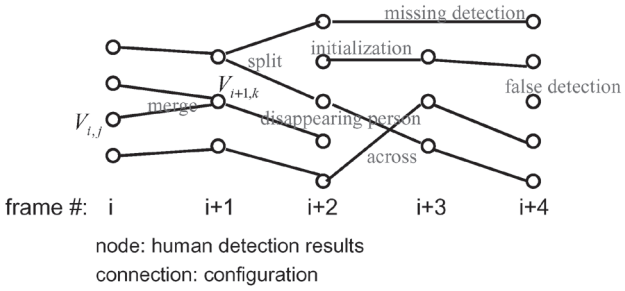

Fig. 2 Graph structure in multiple object tracking.

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 the 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 Fig. 2, the generation process takes care of object occlusions 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 nodes are 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 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 module 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.

5.2 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:

$$\begin{aligned} likelihood_i = {} & likelihood_{i-1} \\ & + \frac{\sum_j^n \left(\log(p_{con_j}) + \log(p_{obj_j}) + \log(p_{trj_j})\right)}{n} \\ & + L_{img} \end{aligned} \tag{2}$$

where i is the current frame number, n represents the number of objects in the current hypothesis. p_{con_j} denotes the connection probability of the j'th trajectory computed in Eq. (1). If the j'th 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 the j'th trajectory. We use the average likelihood of the multiple trajectories for the hypothesis likelihood 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 the following two items:

$$L_{img} = l_{\text{cov}} + l_{comp} \tag{3}$$

where

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

In the above equation, A denotes the sum of the foreground pixels in the current frame, B_j represents the pixels covered by the j'th object in the hypothesis, c is a constant parameter, $\cap$ denotes the set intersection and $\cup$ the set union. l_{cov} computes the hypothesis coverage of the foreground pixels and l_{comp} measures the hypothesis compactness. The higher the l_{cov} value, the larger foreground coverage by the hypothesis, whereas the larger the l_{comp} value, the more compact of the objects making up the hypothesis. These two values give a spatially global explanation of the image (foreground) information. This computation is similar to the image likelihood computation in the reference [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 explanations. It tolerates missing and false detections since it has a global view of image sequences.

5.3 Hypotheses Management

This step ranks the hypotheses according to their likelihood values. To avoid combinatorial explosion in the 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 method 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 pruning 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 pruning 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 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.

6. EXPERIMENTS

The multiple object tracking method was tested on two existing CCTV cameras installed at the main entrance of our lab. The first scenario includes two persons coming into the door at about the same time. **Figure 3**(a) shows four images from the sequence with overlaid bounding boxes showing the human detection results. The darker the bounding box the higher the detection probability. Figure 3(b) demonstrates the multi-tracks with the largest probability generated by the multiple object tracking. 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 the occlusion, and the person in the front due to the 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 Fig. 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 shows the person in white shirt opens the door and lets the person in black shirt into the door without swiping his card. Figure 4(a) shows the images from the sequence and (b) demonstrates the tracking result. Therefore, two tracks are shown in Fig. 4(b), the long track for the person leaving the door and the short track for the man entering the door.

(a) **(b)**

Fig. 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.

(a) **(b)**

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

7. CONCLUSION

In this paper, we presented a novel multi-object tracking technique that integrates object detection and tracking and discovers the globally optimized object trajectories through multiple hypothesis analysis. This tracking technique is the central piece of the SmartCatch intelligent video surveillance system, which has been up and running for more than a year in several major U.S. airports. The SmartCatch system has undergone, and passed the test of various climate and illumination conditions, and its performance has exceeded the customers' expectations in event detection accuracies.

REFERENCES

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

[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*, **38**, 3, pp.791-812, July 2002.

[5] D. B. Reid, "An algorithm for tracking multiple targets," AC, **24**, 6, pp.843-854, Dec. 1979.

[6] T. E. Fortmann, Y. Bar-Shalom and M. Scheffe, "Sonar tracking of multiple targets using joint probabilistic data association," *IEEE J. Oceanic Eng.*, **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 multi-target tracking as an incomplete data problem," *IEEE Trans. on Aerospace and Elec. Sys.*, **33**, 4, pp.1242-1257, Oct 1997.

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

[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, pp.II:194-201, 2001.

[13] C. Stau_er and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *PAMI*, **22**, 8, pp.747-757, Aug. 2000.

Received January 4, 2005

* * * * * * * * * * * * * * * *

Yihong GONG received his B.S., M.S. and Ph.D. degrees in Electronic Engineering from the University of Tokyo in 1987, 1989 and 1992, respectively. He then joined the Nanyang Technological University of Singapore, where he worked as an assistant professor in the School of Electrical and Electronic Engineering for four years. From 1996 to 1998, he worked for the Robotics Institute, Carnegie Mellon University as a project scientist. He was a principal investigator for both the Informedia Digital Video Library project and the Experience-On-Demand project funded in multi-million dollars by NSF, DARPA, NASA and other government agencies. Now he works as a senior research staff for NEC Laboratories America, leading the Rich Media Processing group. His research interests include image and video analysis, multimedia database systems, and machine learning.

* * * * * * * * * * * * * * * *