Observe-and-Explain: A New Approach for Multiple Hypotheses Tracking of Humans and Objects

M. S. Ryoo and J. K. Aggarwal
Computer & Vision Research Center / Department of ECE
The University of Texas at Austin
{mryoo, aggarwaljk}@mail.utexas.edu

Abstract

This paper presents a novel approach for tracking humans and objects under severe occlusion. We introduce a new paradigm for multiple hypotheses tracking, observe-and-explain*, as opposed to the previous paradigm of* hypothesize-and-test*. Our approach efficiently enumerates multiple possibilities of tracking by generating several likely 'explanations' after concatenating a sufficient amount of observations. The computational advantages of our approach over the previous paradigm under severe occlusions are presented. The tracking system is implemented and tested using the* i-Lids *dataset, which consists of videos of humans and objects moving in a London subway station. The experimental results show that our new approach is able to track humans and objects accurately and reliably even when they are completely occluded, illustrating its advantage over previous approaches.*

1. Introduction

Tracking, an automated calculation of object trajectories from video data, is an important problem. The reliable tracking of humans and objects is particularly essential in computer vision for recognition of human activities, which will enable construction of various applications including smart surveillance systems, human-computer interaction systems, and monitoring systems in public places. For example, the tracking of a suitcase and its owner is necessary for the recognition of a human-object interaction of a person abandoning baggage. Also, in order to recognize group activities, a sports play analysis for example, trajectories of players must be obtained. Multiple targets (i.e. humans and objects) must be tracked (MTT), and their trajectories must be analyzed.

However, occlusions among multiple humans, objects, and scene objects make tracking a difficult problem. Figure 1 shows several types of occlusions that may occur in typical monitoring environments: (a) a person occluding other persons, (b) a scene (e.g. a pillar) occluding persons, and (c) the scene occluded by another scene object (e.g. subway train). Tracking a human fully occluded by a scene or another human is a particularly challenging problem, since image information of the person becomes completely unavailable during the period of the occlusion. A tracking system that is able to consider all of the above-mentioned occlusions has not been studied in depth previously.

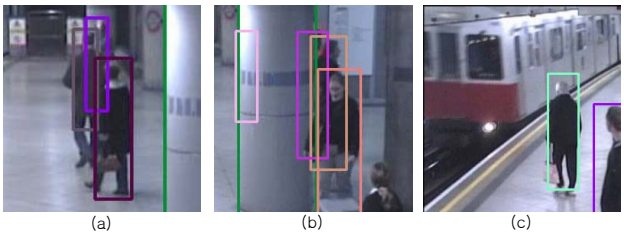

Figure 1. A figure of example occlusions that may occur in monitoring systems.

In this paper, we present a novel methodology for reliable tracking of humans and objects, which is especially designed to handle severe inter-object and scene-object occlusions. We extend the Bayesian formulation of the tracking problem so that it can handle occlusions explicitly, and present a new paradigm for finding optimum trajectories with the maximum a posteriori probability given images. The paradigm of *observe-and-explain* is introduced, as opposed to the previous paradigm of *hypothesize-and-test* that has been widely adopted for probabilistic tracking.

Whenever there are multiple possibilities for tracking, our system goes into 'observation' mode and waits until sufficient information is concatenated, thereby saving unnecessary computations during which the status of the tracking is unclear. Later, in order to enumerate possibilities of tracking, the system probabilistically generates 'explanations' that correspond to observations. The paper describes details of the tracking algorithms following the new paradigm of *observe-and-explain*, while mathematically illustrating

computational advantages of the *observe-and-explain* compared to the *hypothesize-and-test* under severe occlusion. Furthermore, we implement the tracking system following our new paradigm and experimentally verify that the new system performs superior over previous approaches.

In section 2, we discuss several previous tracking approaches. Section 3 describes the Bayesian problem formulation of object tracking. We define the tracking problem as a Bayesian inference of finding the optimum sequence of states. Section 4 presents the paradigm of *observe-and-explain* while comparing it with the previous paradigm of *hypothesize-and-test*. Section 5 illustrates detailed description of implementation of our tracking system following the paradigm of *observe-and-explain*. Experimental results are presented in section 6, and section 7 concludes the paper.

2. Related works

The problem of tracking has widely been studied by previous researchers [13]. Recently, there has been significant effort to address tracking problems under occlusion. Haritaoglu *et al*. [5] designed a head detection method using x and y projections, and successfully segmented a connected foreground region composed of more than two persons. They also proposed a heuristic methodology to track persons overlapped completely. They detected the 'merge' event between two human blobs, and tried to match the persons before the 'merge' and after the 'merge'. Elgammal and Davis [4] attempted to handle occlusions more explicitly, analyzing multiple hypotheses describing which person is in front of the other person when occluded. Zhao and Nevatia [14] estimated the depth information of detected persons based on their head position, and used it to handle occlusions. Wu and Nevatia [12] adopted part-based models to overcome partial occlusions between tracked persons.

Reid [7] has established the concept of a multiple hypothesis tracker (MHT), and proposed a system for probabilistic tracking. Several researchers have conducted research to extend the MHT framework, while focusing on the method to overcome the main drawback of the MHT: MHT requires an exponential amount of computations to enumerate all possible hypotheses. As a result, various approximation algorithms have been developed for efficient tracking. Some of them have limited the total number of hypotheses maintained using pruning [3, 4], while others have developed heuristic tracking algorithms such as those using particle filtering [10, 6] and Markov chain Monte Carlo (MCMC)-based methods [15, 9].

What we must note is that all of the above mentioned approaches with multiple hypotheses derived from [7] follow the paradigm called *hypothesize-and-test*: they generate multiple hypotheses when the state of the tracking is unclear, and evaluate the hypotheses as new observations are given later. However, this paradigm may either take an exponential amount of computations or lead to poor performance when an object (or a person) becomes fully occluded while moving, since there does not exist any clue to prune the hypotheses. A person passing a pillar in a subway station environment is a good example. Linear estimation of movements, using a Kalman filter for example [8], is one solution. However, linear estimations cannot handle movement changes during the occlusion period (e.g. a person changing the direction or velocity while occluded), and thus is not accurate. Therefore, in this paper, we present an alternative paradigm for handling multiple hypotheses called *observe-and-explain*, in order to overcome the problem of *hypothesize-and-test* for handling long-term occlusions.

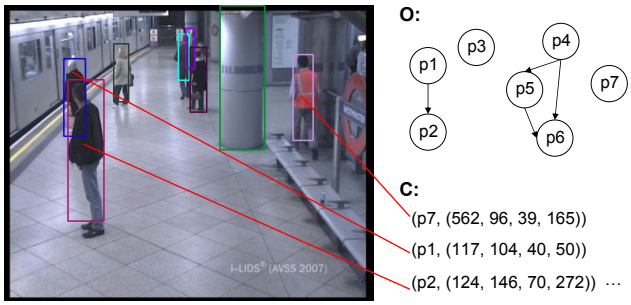

Figure 2. Figure of an example state describing an image frame.

3. Bayesian problem definition

Here, we define the problem of tracking as a Bayesian inference of calculating the posterior probabilities of trajectories. Trajectories of humans and objects in a video are modeled as a sequence of 'states' (or 'tracking states'), where each of them describes the locations of persons (or objects) being tracked in an image frame. In previous systems with a Bayesian tracking problem definition, these locations have commonly been described in terms of 2D coordinates or estimated 3D coordinates [7, 3, 10, 6, 15, 9]. In our approach, a tracking state of each time frame, S_i, is defined as a set of 2D coordinates of bounding boxes of persons, C_i, together with a relative depth order among persons overlapped, O_i: $S_i = (C_i, O_i)$ at time frame $1 \leq i \leq n$. More specifically, C_i is a union of a person's id associated with the (x, y) coordinate and (width, height) of its bounding box, $C_i = \cup\{(k, (x_i^k, y_i^k, w_i^k, h_i^k))\}$, and O_i is a directed graph describing the order: a 'child' person is occluding its 'parent'. Figure 2 shows an example state of an image frame. In summary,

$$S_i = (C_i, O_i)$$

where $C_i = \cup\{(k, (x_i^k, y_i^k, w_i^k, h_i^k))\}$, and $O_i =$ a directed graph describing the order among persons.

As a result, trajectories of persons in a sequence of image frames can be described as a sequence of states: $S_1, S_2, S_3, ..., S_n$. In the Bayesian formulation of track-

ing, the goal is to find the optimal sequence of states that maximizes the posterior probability of the state sequence given the input frames. That is, the goal is to detect $argmax_{(S_1,S_2,...,S_n)} P(S_1, S_2, ..., S_n | I_1, I_2, ..., I_n)$ where each I_i indicates an image at frame i.

In principle, there are an exponential number of possible sequences to enumerate: $O(r^n)$ where r is an average number of possible states per single image frame. Further, the size of r is quadratic (or even larger) to the size of the image frame, preventing a brute force search from being applied.

4. *Hypothesize-and-test* vs. *observe-and-explain*

4.1. *Hypothesize-and-test*

Hypothesize-and-test is a paradigm for heuristic methodologies to find a solution for the Bayesian formulation of the tracking problem. In the *hypothesize-and-test* paradigm, each sequence of states $(S_1, S_2, S_3, ..., S_n)$ is treated as a 'hypothesis' of the tracking, and the system 'tests' them by calculating a posteriori probability for each hypothesis. In order to estimate the optimum hypothesis among the exponential number of possible hypotheses, *hypothesize-and-test* approaches take advantage of other existing elementary object detectors and trackers. Instead of searching for all possibilities for each frame, *hypothesize-and-test* approaches maintain a few promising 'hypotheses' and update (and diverge) each hypothesis using an elementary detector and tracker. Object detection methods such as a foreground blob detector using a background subtraction have widely been used as elementary detectors, while tracking algorithms like meanshift and blob trackers have been adopted as elementary trackers.

In approaches following the *hypothesize-and-test* paradigm, whenever a new image frame at time n is given, each hypothesis of frame 1 to frame $(n-1)$, $(S_1, S_2, S_3, ..., S_{n-1})$ is updated to obtain $(S_1, S_2, S_3, ..., S_n)$ by calculating the new state S_n based on the state S_{n-1}. Results of elementary detectors and trackers applied to the frame n are used to update the state S_{n-1}. That is, Markov assumptions are usually made so that an elementary tracker can efficiently update each hypothesis iteratively. This update may cause one hypothesis to diverge into two or more hypotheses, if there are more than two possibilities of the status change of the person. For example, a hypothesis may diverge into two if there exists a newly detected person: one including the new person to be present, and the other excluding the new person as a noise detection.

Maintained hypotheses are pruned probabilistically by 'testing' the hypothesis. After updating hypothesis at time n, the *hypothesize-and-test* system evaluates each of them by measuring the posterior probability of the hypothesis $(S_1, S_2, S_3, ..., S_n)$ given a sequence of images $(I_1, I_2, I_3, ..., I_n)$. More formally, the probability of updated hypotheses can be calculated as follows:

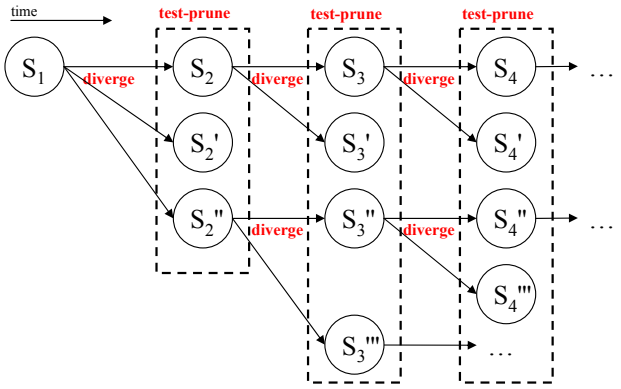

Figure 3. An example process of a *hypothesize-and-test* system.

$$argmax_{(S_1,...,S_n)} \; P(S_1, ..., S_n | I_1, ..., I_n) = argmax \; E(n)$$

where

$$\begin{aligned} E(n) &= P(I_1, ..., I_n | S_1, ..., S_n) \cdot P(S_1, ..., S_n) \\ &= P(I_n | S_n) \cdot P(S_n | S_{n-1}) \\ &\quad \cdot P(I_1, ..., I_{n-1} | S_1, ..., S_{n-1}) \cdot P(S_1, ..., S_{n-1}) \\ &= P(I_n | S_n) \cdot P(S_n | S_{n-1}) \cdot E(n-1) \end{aligned}$$

In most systems, rules to make a decision to diverge a hypothesis are encoded a priori for computational efficiency of the system. For example, as mentioned above, a rule specifying that the system should diverge a hypothesis when a new person is likely to be added or removed from the scene must be encoded a priori. Figure 3 illustrates the process of the *hypothesize-and-test* system.

Diverging processes may lead the system to maintain an exponential number of hypotheses. Most of the previous systems limited the number of maximum hypotheses to maintain, in order to prevent this problem. Approaches using particle filtering or Markov chain Monte Carlo (MCMC) avoid the hypothesis diverging process by performing stronger hypothesis analysis at each frame (i.e. pruning all but the most promising one) with probabilistic reversible dynamics. However, in the case of full occlusions, strong pruning may be dangerous since the appearance of occluded persons (or objects) becomes unobservable, preventing any meaningful evaluation on hypotheses.

4.2. *Observe-and-explain*

We introduce our new paradigm called *observe-and-explain*, an alternative approach for finding the sequence of tracking states that maximizes the posterior probability. *Observe-and-explain* enables the efficient enumeration of multiple possibilities of tracking, thereby improving reliability and accuracy of the tracking system. Instead

of diverging and maintaining all intermediate possibilities at every time frame, *observe-and-explain* 'observes' until enough information is concatenated to make any meaningful analysis, and then probabilistically generates the most likely 'explanations' on the movements of a person corresponding to the observation history.

The intuition behind the *observe-and-explain* is to enumerate tracking possibilities only when the system has enough information to evaluate them. As a result, the *observe-and-explain* approach avoids the enumeration of an exponential number of possibilities that may occur when pruning is not possible due to insufficient data (e.g. during when a person is fully occluded). If the system deduces that there may be multiple possibilities regarding the status of a person, the system postpones analyzing the status of the person temporarily until the system 'observes' (i.e. receives) future image frames. Later, after the system decides that it has sufficient information to analyze the status of the person, multiple 'explanations' are generated stochastically. An 'explanation' is a candidate sequence of states that is likely to match input frames given during the 'observation' period. The detailed process of the approach is as follows.

If there is no occlusion, appearance, or disappearance of persons, the system iteratively updates the locations of persons being tracked using an elementary detector and tracker. That is, if state diverging is not necessary, the system always has a sufficient amount of information to update the tracking state of all persons: a single image of the next frame is sufficient to update the state using an elementary tracker. However, when the system decides that the tracking process must consider multiple possibilities because of a location of a particular person (or because of a relationship between two persons), the system labels the location of the person (or the relationship between the persons) as 'to be determined' and continues updating the other persons. The person (or the relationship between persons) is left as 'to be determined' until the system deduces that sufficient information to analyze possibilities of the status of the person (or the relationship between persons) has been obtained. After the system obtains sufficient information, the system generates several candidate 'explanations', which is a sequence of the status of the person labeled as 'to be determined' in the observed image frames. The likelihood between explanations and image frames observed are measured probabilistically, and the ones with low probability are pruned.

Figure 4 illustrates the process of the *observe-and-explain* system. The system may prune all but the state with the most likely 'explanation' or may maintain multiple hypotheses stochastically, depending on the implementation. Similar to a *hypothesize-and-test* approach, rules to label a location or a relationship as 'to be determined' and rules to remove the label by generating 'explanations' are generally encoded a priori, for efficient tracking.

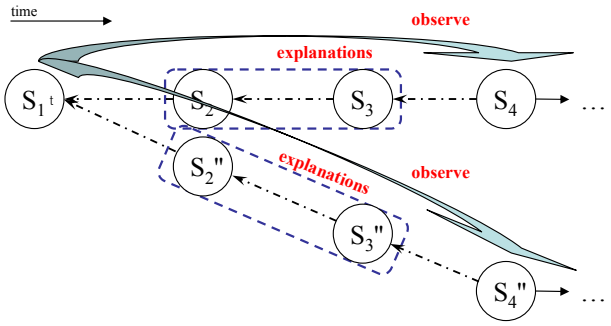

Figure 4. An example process of an *observe-and-explain* system.

Formal equations for labeling processes and label removing processes are as follows:

$$argmax_{(S_1,...,S_n)}\ P(S_1,...,S_n|I_1,...,I_n) = argmax\ F(S_1,...,S_n)$$

For update:

$$F(S_1,...,S_n) = F(S_1,...,S_{n-1}) \cdot P(I_n|S_n) \cdot P(S_n|S_{n-1})$$

For going into the observation period:

$$F(S_1,...,S_n^{-k}) = F(S_1,...,S_{n-1}) \cdot P(I_n|S_n^{-k}) \cdot P(S_n^{-k}|S_{n-1})$$

Update during observation period:

$$F(S_1,...,S_n^{-k}) = F(S_1,...,S_{n-1}^{-k}) \cdot P(I_n|S_n^{-k}) \cdot P(S_n^{-k}|S_{n-1}^{-k})$$

For generating explanations to exit observation period:

$$F(S_1,...,S_n) = F(S_1,...,S_{m-1},S_m^{-k},...,S_n^{-k}) \cdot P(s_m^k,...,s_n^k) \cdot \prod_{i=m\ to\ n} P(I_i^k|s_i^k)$$

where S_n^{-k} indicates kth object in the scene is labeled as 'to be determined', I_i^k is an image region related to the kth object in the scene, and s_i^k is state information related to the kth object in the scene S_i. The labeling of multiple persons as 'to be determined' is also possible. Multiple labels are concatenated and maintained independently.

The main advantage of the *observe-and-explain* is on its ability to efficiently track objects even if they are fully occluded. Assume that a person walked behind a pillar while the system is tracking him/her, for example. During the period of the occlusion, the *observe-and-explain* system just 'observes' given images, even though there are an exponential amount of possibilities of movements of the person behind the pillar. Later, if a new person is detected around the pillar, the system will generate a limited number of 'explanations': one describing the person is still behind the pillar, and the others describing that the person who went behind the pillar just came out. An 'explanation' assuming linear motion of the person will have the highest prior probability. As a result, the *observe-and-explain* is able to save $O(q^t)$ amount of unnecessary computations during the occlusion period, where q is an average number of motion possibilities at each frame, and t is the period of the occlusion.

5. Implementation of tracking system with *observe-and-explain*

In this section, we present a detailed algorithm and an implementation of the tracking system for humans and objects based on *observe-and-explain*. We first describe the elementary detectors and trackers which our new tracking system takes advantage of. A method to calculate the likelihood between an image and a state is presented next, which is essential for a posteriori calculation process. Finally, heuristic rules for updating the state, rules to label a tracked human in order to enter 'observation period', and rules to remove labels by constructing 'explanations' are presented. These rules enable an efficient search of the optimum tracking solution with the *observe-and-explain*.

5.1. Elementary detectors and trackers

In order to provide information on a person's existence, two elementary object detectors have been implemented. One is a human-blob segmentation method using a background subtraction technique. Similar to [5] and [14], foreground blobs are segmented, and peaks of foreground regions are detected to estimate positions of human heads. Peaks which are reliably detected throughout a certain period of time frames are considered candidate head positions of a person. The second elementary detector is head detection using Viola and Jone's object detector [11]. As a result, bounding boxes of humans are estimated based on the locations of peaks and heads using the 3D camera model of the scene. At each frame, results of the elementary detectors are given to the high-level of the system, so that it can update the state of the scene based on the detections.

In addition, we implement an elementary tracking algorithm in order to help the efficient updating of states. Each person who is currently in the state is tracked using an elementary tracker. We implement a tracker adopted from [14]. The tracker maintains an appearance model of each person in the scene. The appearance model consists of two arrays whose sizes are normalized. One array, M, contains a mean pixel value of a person, and the other array, W, contains a probability of each pixel being a foreground. Figure 5 shows an example of an appearance model of a person.

The elementary tracker estimates the next location of the person by matching the image frame with the appearance model of the person. The coordinate information of the persons, C_{n-1}, and their relative depth order, O_{n-1}, are used to search for the next location of the person. Basically, the location of kth person at frame n, c_n^k, is first estimated based on the location of c_{n-1}^k and the velocity of the person k. The system searches for nearby image regions from the next location of the person estimated using its velocity with the minimum distance between an image region and the appearance model. When matching, the perspective distortion is considered to update the size changes of bounding boxes.

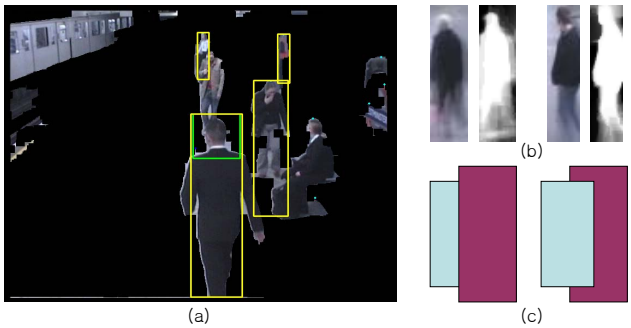

Figure 5. (a) shows detection results of the elementary detectors. Green rectangles are the heads detected and cyan dots are the peaks. Yellow rectangles are the estimated bounding boxes based on the elementary detectors. (b) shows example appearance models of persons being tracked. (c) shows an example order of the matching-masking task of an elementary tracker.

The matching process is done starting from the person who is closest to the camera, and continues following the relative depth order, which is specified in O_n. Once an image region is decided to be corresponding to a person, the image region is masked so that it does not influence the matching process of the persons on his/her back. Formally,

$$(x_n^k, y_n^k, w_n^k, h_n^k) = argmax_{(x,y,w,h)} \sum_p (match(M_p^k, p) \cdot W_p^k)$$

where p is an each pixel in (x, y, w, h), and $match(M_p^k, p)$ is a constant value if p is already masked.

5.2. Measuring image likelihoods

Assume that a state describing the persons in the current scene is provided. In order to evaluate an explanation and calculate a posteriori probabilities, the system must be able to calculate the likelihood of the image, $P(I_i|S_i)$. We assume conditional independence among persons in the image given the state. That is, we calculate $P(I_i|S_i)$ as:

$$P(I_i|S_i) = \prod_{k=0\ to\ max_k} P(I_i^k|s_i^k) = \prod_{k=0\ to\ max_k} P(I_i^k|b_i^k)$$

where b_i^k indicates a bounding box region of kth object not occluded by others.

The likelihood between each image region and a given bounding box from the state, $P(I_i^k|b_i^k)$, is measured by counting a ratio of a foreground region and by calculating pixel-wise color distances. That is,

$$P(I_i^k|b_i^k) = P(FgrndLkhd_i^k|b_i^k) \cdot P(ColorLkhd_i^k|b_i^k)$$

where we estimate $P(FgrndLkhd_i^k|b_i^k)$ to have a Gaussian distribution over the ratio of foreground pixels in the region b_i^k, and $P(ColorLkhd_i^k|b_i^k)$ to have a Gaussian distribution over the sum of color distances of pixels in b_i^k.

In addition, we consider the probability of a state indicating a person 'not existing' in a scene. Since handling 'addition' and 'deletion' of a person from the state will stochastically generate two possibilities (one with a person and the other without a person), the system must have the ability to calculate the probability of a person in 'nowhere'. Similar to positive likelihood calculations, our system models the likelihood of a person 'not' in a certain location b_i^k, $P(I_i^k|\neg b_i^k)$, in terms of two independent Gaussian distributions: $P(BkgrndLkhd_i^k|\neg b_i^k)$ and $P(ColorNotLkhd_i^k|\neg b_i^k)$. Since an identical person can exist at only one location at the same time, we calculate the expectation of the probability of a kth person being nowhere, $P(I_i^k|\neg s_i^k)$, by calculating $P(I_i^k|\neg b_i^k)$ based on the locations that the system believes the object to be.

If a person k is labeled as 'to be determined' because of the observation period, the system sets $P(I_i^k|s_i^k) = 1$.

5.3. State update policies for *observe-and-explain*

We have discussed fundamental equations in section 4.2. The rules present in this subsection are an implementation of *observe-and-explain*, which has been designed to track humans and objects under severe occlusions such as persons moving while fully occluded. At each frame, the system needs to update the state sequence of the currently maintained hypotheses based on results from elementary detectors and trackers. Probabilities of state transition as well as likelihoods we discussed in the previous subsection are considered for a posteriori probability calculation.

Basic updates. If there is no more than one possibility of an update, the system simply updates the state based on the results from the elementary tracker. We assume that the results of the elementary tracker are the optimum updates that maximize a posteriori probability. As a result, S_{n-1} is updated to obtain S_n as follows.

$$F(S_1, ..., S_n) = F(S_1, ..., S_{n-1}) \cdot P(I_n|S_n) \cdot P(S_n|S_{n-1})$$

'Explaining' additions and deletions. When one of the elementary detectors detects a new person from the next image frame, there are three possibilities of updates: the update considering new detection as a new location of an existing person, the update treating new detection as a newly appearing person, and the update ignoring the new detection as noise. Whenever the new detections are provided, three 'explanations' corresponding to three cases must be generated and evaluated. In order to construct explanations, the newly detected person is tracked reversely starting from the current frame, for a certain number of previous frames. The intuition is to construct history trajectories of the newly detected person to analyze whether the new detection is noise.

$$F(S_1, ..., S_{n-1}, (C'_n, O'_n)) = F(S_1, ..., S_{n-1}, (C_n, O_n)) \cdot a1$$

$$F(S_1, ..., S_{n-t-1}, (C_{n-t} + c^k_{n-t}, O_{n-t}), ..., (C_n + c^k_n, O_n))$$
$$= F(S_1, ..., S_n) \cdot a2 \cdot \prod_{i=n-t \text{ to } n} P(I_i^k|s_i^k)$$

$$F(S_1, ..., S_{n-t-1}, S_{n-t} + \neg s^k_{n-t}, ..., S_n + \neg s^k_n)$$
$$= F(S_1, ..., S_n) \cdot (1 - a1 - a2) \cdot \prod_{i=n-t \text{ to } n} P(I_i^k|\neg s_i^k)$$

where C_n is coordinates of tracked persons which are updated based on the elementary tracker, C'_n is coordinates of tracked persons which are updated based on the new detection, c^k is a coordinate of the new kth person, t is a number of time frames to analyze, and $+$ sign indicates that we are adding new coordinate information to an existing state.

Constant $a1$ is the a priori probability of an elementary detector detecting a new location of an existing person, and $a2$ is the a priori probability of a new person appearing in a scene. $a2$ generally is significantly smaller than $(1 - a2)$. In our implementation, we have empirically chosen $a1$ to be 0.3 and $a2$ to be 0.1.

Equations for handling deletions of existing bounding boxes of tracked persons can be posed in a similar fashion:

$$F(S_1, ..., S_{n-1}, (C_{n-t} - c^k_{n-t} + \neg c^k_{n-t}, O_{n-t}),$$
$$..., (C_n - c^k_n + \neg c^k_n, O_n))$$
$$= F(S_1, ..., S_n) \cdot a3 \cdot \prod_{i=n-t \text{ to } n} \Big(P(I_i^k|\neg s_i^k)/P(I_i^k|s_i^k)\Big)$$

where c^k is a coordinate of the existing kth person and $a3$ is a priori probability of a removal. We empirically set $a3$ as 0.1 when it is not near the scene boundary and as 0.9 when near the boundary.

'Explaining' occlusions. There are two possibilities of a depth order between two occluded persons, and the tracking system must find the order which better matches with the observed images. However, at the initial stage of occlusion (i.e. when two persons are simply touching), there is not enough information to analyze the ordering between the two. Therefore, the system must go into the observation mode.

If $overlap(k, l)$ becomes larger than 0 at time n,

$$F(S_1, ..., S_n^{-k,-l}) = F(S1, ..., S_{n-1}) \cdot P(S_n^{-k,-l}|S_{n-1})$$

Later, when the ratio of an occluded area exceeds a certain threshold, the system generates two 'explanations' describing the relative depth order.

If $overlap(k, l)$ becomes larger than τ at time n,

$$F(S_1, ..., (C_n, O_n^{k>l})) = F(S_1, ..., S_{m-1}, S_m^{-k,-l}, ..., S_n^{-k,-l})$$
$$\cdot d \cdot \prod_{i=m}^{n} P(I_i^k|(C_i, O_i^{k>l})^k) \cdot \prod_{i=m}^{n} P(I_i^l|(C_i, O_i^{k>l})^l)$$

$$F(S_1, ..., (C_n, O_n^{k<l})) = F(S_1, ..., S_{m-1}, S_m^{-k,-l}, ..., S_n^{-k,-l})$$
$$\cdot (1 - d) \cdot \prod_{i=m}^{n} P(I_i^k|(C_i, O_i^{k<l})^k) \cdot \prod_{i=m}^{n} P(I_i^l|(C_i, O_i^{k<l})^l)$$

where $O_i^{k>l}$ indicates the kth object is in front of lth object in frame i. The variable d is a prior probability, which we model to have a logistic distribution over $(y_n^k - y_n^l)$, whose mean is set to 0.

'Explaining' occluded motion. When a person becomes fully occluded by a larger object (e.g. a pillar), the system must stop tracking the person and go into the observation mode. That is,
If $overlap(k, l)$ becomes almost 1 at time n and k is the occluded object,

$$F(S1, ..., S_n^{-k}) = F(S1, ..., S_{n-1}) \cdot P(S_n^{-k}|S_{n-1})$$

Whenever a new person appears (i.e. whenever an elementary detector detects a new person) near the object occluding the person, there are two possibilities. The newly detected person might be an irrelevant person, or he/she might be the person who hid behind the pillar previously. In the former case, the person who hid must still be behind the pillar and the newly detected person must be handled as we have discussed in the subsection "Explaining addition and deletion". In the latter case, the person must have moved through the pillar during the observation period.

We synthesize the most likely motion of the person who hid, assuming that the person's latest position is the location of the person newly detected. The motion we generate is a two-step linear motion: the first part is the person going into the object occluding the person, and the second part is the person going out of the occluding object. Movements are estimated based on the velocities of the person before being occluded and the person after reappearing. Based on the motion estimated, the locations of the person during the observation period, s_i^k, can be recovered, enabling the calculation of $P(I_i^k|s_i^k)$. Thus, the probability function F is:

$$F(S_1, ..., S_n) = F(S_1, ..., S_{m-1}, S_m^{-k}, ..., S_n^{-k}) \cdot e \cdot \prod_{i=m \ to \ n} P(I_i^k|s_i^k)$$

$$F(S_1, ..., S_n^{-k}) = F(S_1, ..., S_{n-1}^{-k}) \cdot (1 - e) \cdot P(I_n|S_n^{-k})$$

where the variable e is a prior probability associated with the motion of the occluded person. We estimate the variable e to have a Gaussian distribution over difference between velocities of two steps of the motion, whose mean is 0.

6. Experiments

We test the system on the *i-Lids* dataset [2]. The *i-Lids* dataset is composed of videos taken at a subway station in London, UK at the resolution of 720*576 at the rate of 25 fps. The video contains a large number of occlusion events among pedestrians and a pillar. Total of 5060 frames of the *AVSS-AB-hard* sequence have been used as test data. We not only implement our new system following the paradigm of *observe-and-explain*, but also construct a system following the previous approach of *hypothesize-and-test* which can be viewed as our implementation of [7, 3]. Two systems are implemented to spend a similar amount of computations, by making *hypothesize-and-test* to maintain maximum h hypotheses. The tracking accuracy of the systems are measured and compared. In both systems, two likelihood functions ($P(FgrndLkhd_n^i|b_n^i)$ and $P(ColorLkhd_n^i|b_n^i)$) and priors are estimated with separate training data. Locations of scene objects (e.g pillar) and scene boundaries are provided to the systems a priori.

System	Total	TP	Fgmt	Drft	Swtch	FP
ObsvAndExpn	72	49	15	11	1	4
HypoAndTest	72	33	27	13	3	4

Table 1. Overall tracking accuracy of the systems on *i-Lids* dataset.

System	Full SO	Full OO	Partial OO	Total
ObsvAndExpn	35/45	18/31	29/34	82/110
HypoAndTest	16/45	14/31	27/34	57/110

Table 2. Accuracy of the systems in handling occlusions.

Table 1 shows the overall tracking accuracy of two systems on the test dataset. TP measures true positive trajectories (i.e. totally correct) without any errors, and FP shows the number of false positive persons the system detected. $Fgmt$, $Drft$, and $Swtch$ are types of tracking errors, where $Fgmt$ indicates that a trajectory obtained has been fragmented, $Drft$ indicates that a tracked person's ID has been drifted to another person, and $Swtch$ specifies the number of ID switches between two different persons. In addition, the performance of the system in handling occlusions (e.g. whether the system was able to successfully track a person moving behind a pillar) is measured in table 2. SO indicates "scene occlusions" and OO indicates "inter-object occlusions". We are able to observe that our system with *observe-and-explain* performs superior to the previous system, mainly because of our system's ability to handle full occlusions. As we can see from table 2, *observe-and-explain* handles full occlusions (both scene-object and inter-object) more reliably than the previous system, thus causing the overall accuracy shown in table 1 to increase.

In addition, in order to illustrate the robustness of our system, we analyze the difficulty of the *i-Lids* dataset over other datasets in the aspect of occlusions. We compare the rate of an average person being occluded (at least partially) and fully occluded in two datasets: the *i-Lids* dataset and the *CAVIAR* dataset [1], which has been commonly used for evaluating a pedestrian tracking system. In the case of the *CAVIAR* dataset, an average number of occlusions per person is approximately 0.5 and that of full occlusions is 0.15. In the case of *i-Lids* dataset, the numbers are 1.52 for

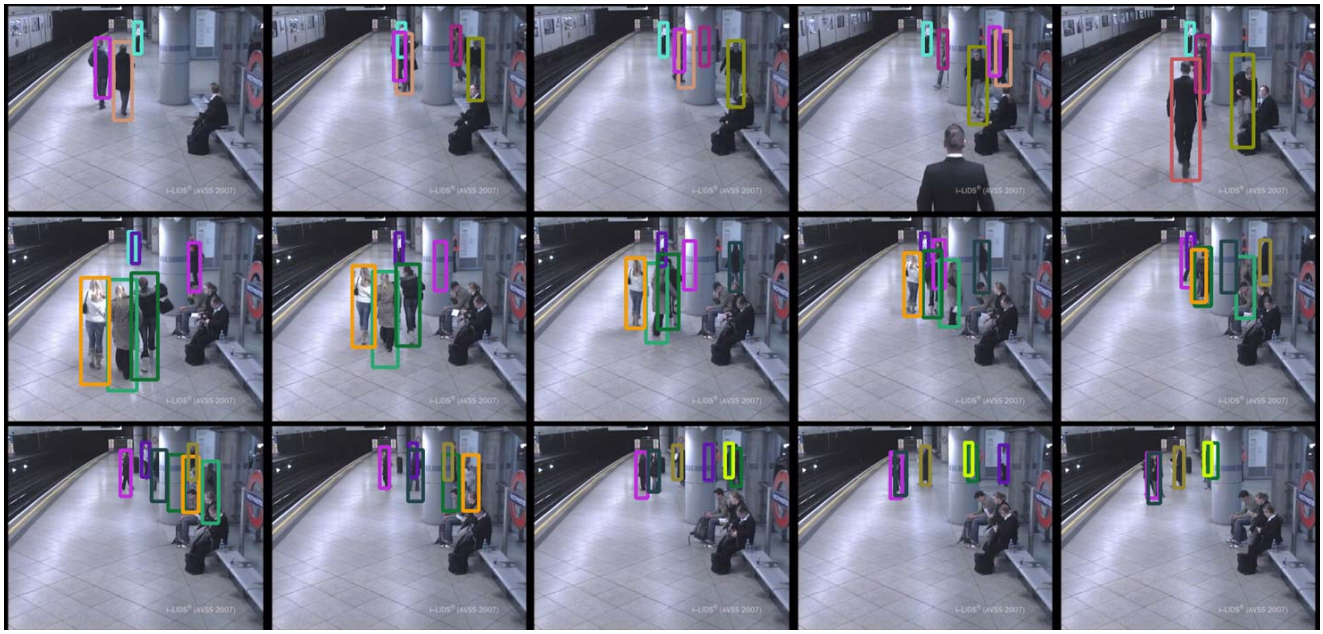

Figure 6. Example tracking results of our system tested on the *i-Lids* dataset. (This picture is best viewed in color)

occlusions and 1.05 for full occlusions. In the *i-Lids* dataset, full occlusion occurs frequently because of the pillars.

The state-of-the-art systems show 0.6˜0.7 accuracy on the *CAVIAR* dataset [15, 12]. We are able to observe that our new system performs comparably to previous systems (or better than in the aspect of occlusion handling) even with more difficult dataset, *i-Lids*. Furthermore, we have tested half of the *CAVIAR* dataset (all files with suffix *"1cor.mpg"*), obtaining overall tracking accuracy of 0.78.

7. Conclusions

We have presented the paradigm of *observe-and-explain* for the tracking of humans and objects. Our *observe-and-explain* is able to enumerate multiple possibilities of tracking efficiently, thereby enabling robust and reliable tracking even under severe occlusions. We designed and implemented a tracking system following our new paradigm, and have verified the advantages of our paradigm through conducting experiments with the dataset which contains severe occlusions. In future, we plan to further exploit tracking approaches with the paradigm of *observe-and-explain*, by incorporating stronger elementary detectors and trackers and by applying other search techniques to enhance the system.

References

[1] http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1/.

[2] i-lids dataset for abandoned baggage scenario. http://www.elec.qmul.ac.uk/staffinfo/andrea/avss2007_d.html.

[3] I. J. Cox and S. L. Hingorani. An efficient implementation of Reid's multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking. *IEEE T PAMI*, 18(2):138–150, 1996.

[4] A. M. Elgammal and L. S. Davis. Probabilistic framework for segmenting people under occlusion. In *ICCV*, pages 145–152, 2001.

[5] I. Haritaoglu, D. Harwood, and L. S. Davis. W4: Real-time surveillance of people and their activities. *IEEE T PAMI*, 22(8):809–830, 2000.

[6] M. Isard and J. MacCormick. Bramble: A Bayesian multiple-blob tracker. In *ICCV (2)*, page 34, 2001.

[7] D. Reid. An algorithm for tracking multiple targets. *IEEE T Automatic Control*, 24(6):843–854, Dec 1979.

[8] R. Rosales and S. Sclaroff. 3d trajectory recovery for tracking multiple objects and trajectory guided recognition of actions. In *CVPR*, pages 2117–2123, 1999.

[9] K. Smith, D. Gatica-Perez, and J.-M. Odobez. Using particles to track varying numbers of interacting people. In *CVPR (1)*, pages 962–969, 2005.

[10] H. Tao, H. S. Sawhney, and R. Kumar. A sampling algorithm for tracking multiple objects. In *Workshop on Vision Algorithms*, pages 53–68, 1999.

[11] P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR (1)*, pages 511–518, 2001.

[12] B. Wu and R. Nevatia. Tracking of multiple, partially occluded humans based on static body part detection. In *CVPR (1)*, pages 951–958, 2006.

[13] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM Computing Surveys*, 38(4):13, 2006.

[14] T. Zhao and R. Nevatia. Tracking multiple humans in complex situations. *IEEE T PAMI*, 26(9):1208–1221, 2004.

[15] T. Zhao and R. Nevatia. Tracking multiple humans in crowded environment. In *CVPR (2)*, pages 406–413, 2004.