

# Multiple Objects Tracking with Multiple Hypotheses Graph Representation

Alex Yong Sang Chia, Weimin Huang and Liyuan Li  
Institute for Infocomm Research  
21 Heng Mui Keng Terrace, Singapore 119613

## Abstract

*We present a novel multi-object tracking algorithm based on multiple hypotheses about the trajectories of the objects. Our work is inspired by Reid's multiple hypothesis tracking algorithm which is an optimal solution to the motion correspondence that occurs in multi-object tracking. Unfortunately, the exponential growth of the hypotheses tree precludes practical applications. To restrict this growth, many approximations relying on a series of clustering and pruning operations have been proposed. The decisions for these operations are based solely on previous observations and are not guided by observations in later frames. We show that due to multiple splits and merges, relying solely on previous observations to guide these operations may inadvertently eliminate the correct hypothesis. Consequently, this leads to poor tracking performance. To overcome this problem, we determine the validity of a hypothesis by exploiting information in later frames and relating them to previous observations. Experimental results demonstrate the robustness and efficiency of our approach.*

## 1. Introduction

Multi-object tracking (MOT) is a major research topic in computer vision. The main difficulty lies in data association of the measurements to the appropriate tracks. This is further compounded by the presence of occlusions and multiple splits and merges in MOT. Thus, we need an efficient representation to integrate the spatial and temporal information. Such a representation will then enable us to establish patterns and relationships among detected moving regions reliably. Graph representation is one such representation and has attracted interest from the research community [1, 2].

However, these tracking algorithms do not track objects well under severe occlusions and multiple splits and merges. For example, the authors in [1] point out that their algorithm can only yield part of the objects' trajectories. We also note in [2] that their algorithm is unable to track objects robustly through splits and merges. The underlying reason for the

failure of these algorithms is that they keep only *one* hypothesis of the tracking results. Thus they are unable to deal robustly with the ambiguities that arise during the tracking process. In this aspect, multiple hypothesis tracking algorithms (MHT) may yield better tracking results.

One of the best known MHT is that developed by Reid [3]. Unfortunately, the hypotheses tree grows exponentially as more measurements are received. Many approximations have been proposed to limit this growth by a series of clustering and pruning operations [4, 5]. However, these approximations share a common weakness in that the decisions for these operations are based *solely* on previous observations and are not guided by later frames. We show that previous observations may not provide adequate information to guide these operations in MOT. Consequently, the errors from these wrong operations will propagate into later frames leading to poor tracking performance. M. Han et al. propose another MHT in [6]. However due to high computation complexity, they impose an upper bound on the number of simultaneous active tracks. Furthermore, they do not explicitly deal with the problem of splits and merges. Hence their algorithm is unlikely to work well in a crowded scene.

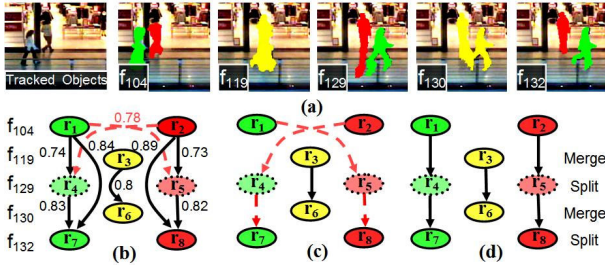
To overcome the drawbacks of these algorithms, we propose a novel multi-object tracking algorithm based on multiple hypotheses about the trajectories of the objects. These hypotheses are generated from the temporal and spatial information of the entire video sequence. We determine the validity of a hypothesis by exploiting information in later frames and relating them to previous observations. Our experiments demonstrate the robustness of our tracking algorithm in which we recover the trajectories of objects under severe occlusions, variations of the lighting condition and multiple splits and merges in camera settings having limited perspective distortion. The limiting factor of our algorithm is that we require the entire video sequence to be available first before we process the video sequence as a whole.

## 2. Introductory Example

Consider the two persons,  $P_a$  and  $P_b$ , shown in the first image of Fig. 1a. The moving regions associated with  $P_a$

and  $P_b$  across five frames are shown in the remaining five images of Fig. 1a. We denote the set of regions in  $f_{104}$ ,  $f_{129}$  and  $f_{132}$  which uniquely reference  $P_a$  as  $\{r_1, r_4, r_7\}$  and  $P_b$  as  $\{r_2, r_5, r_8\}$ . Trajectories of  $P_a$  and  $P_b$  merged in  $f_{119}$  and  $f_{130}$  are detected as  $r_3$  and  $r_6$  respectively. Regions across frames are connected by a directed weighted edge, where the edge weight represents the similarity score between the connected regions (equ. (1)). We represent correct and wrong connections as solid and dashed edges respectively and remove the edges whose weights are too low (Fig. 1b).

Due to the variations of the lighting condition, pixels in the shadows of  $P_a$  and  $P_b$  are detected as foreground pixels following the split in  $f_{129}$ . Thus  $r_4$  and  $r_5$  are wrongly detected as having closer similarity to  $r_2$  and  $r_1$  respectively. At  $f_{132}$ , the distortion from the shadows is reduced. Hence there exist closer similarity between  $r_7$  with  $\{r_1, r_4\}$  and  $r_8$  with  $\{r_2, r_5\}$ . A MHT which makes clustering or pruning decisions based *only* on previous observations may cluster  $r_5$  to  $r_1$  and  $r_4$  to  $r_2$  at current frame  $f_{129}$ . This leads to the wrong tracking results (Fig. 1c). However by exploiting information in later frame  $f_{132}$ , we can eliminate these false connections. Consequently, with the correct connections, we can extract the true trajectories of  $P_a$  and  $P_b$  (Fig. 1d).



**Figure 1.** a) Detected moving regions. b) Hypotheses generated. c) Wrong trajectories extracted due to distortion from the shadows. d) Correct trajectories extracted using information in later frame  $f_{132}$ .

### 3. Graph Representation of Motion

The trajectories of multiple objects can be represented by a set of graphs where the linkages of the moving regions across frames indicate the motion of the objects. These moving regions can be detected using background change detection methods, where adaptive dynamic background modeling approaches give better segmented foreground results. In this paper, we adopt the method developed in [7] to detect these moving regions.

Once the moving regions are extracted, we can determine the trajectories of the objects by connecting the regions which are associated with the same object together

(Fig. 1d). Let  $f_y$  be the  $y^{\text{th}}$  frame of the video sequence,  $1 \leq y \leq n$ , where  $n$  is the available number of frames in the video sequence. We denote the set of moving regions in  $f_y$  as  $R_y = \{r_y^a\}_{a=1}^{m_y}$ , where  $r_y^a$  is a moving region in  $f_y$  and  $m_y$  is the number of moving regions in  $f_y$ . Each moving region,  $r_y^a$ , is also represented as a vertex,  $v_y^a$ , in the graph:  $v_y^a = [p_y^a \quad s_y^a \quad h_y^a]^T$  where  $p_y^a$  is the 2-D position,  $s_y^a$  is the size and  $h_y^a$  is the normalized color histogram of  $r_y^a$ .

We define the relation between two vertices,  $v_y^a$  and  $v_z^b$ , as  $\zeta(v_y^a, v_z^b) = [\zeta_p \quad \zeta_s \quad \zeta_h]^T$ , where  $\zeta_p = |p_y^a - p_z^b|$  is the Euclidean distance,  $\zeta_s = \frac{\max(s_y^a, s_z^b)}{\min(s_y^a, s_z^b)}$  is the scale difference and  $\zeta_h = H(h_y^a, h_z^b)$  is the normalized histogram intersection of the two regions,

$$H(p, q) = 1 - \sum_{j=1}^K |p_j - q_j| \quad (1)$$

where  $\{p_j\}_{j=1}^K = h_y^a$ ,  $\{q_j\}_{j=1}^K = h_z^b$  and  $K$  is the number of bins in the histogram. We represent the directed edge connecting  $v_y^a$  to  $v_z^b$  as  $e_{y,z}^{a,b}$ . The directed edge  $e_{y,z}^{a,b}$  is a logical variable in which  $e_{y,z}^{a,b} = 1$  if  $v_y^a$  is connected to  $v_z^b$  and  $e_{y,z}^{a,b} = 0$  otherwise. We define the set of directed edges arising from and directed to  $v_y^a$  as  $F(v_y^a)$  and  $T(v_y^a)$  respectively.

### 4. Tracking of Multiple Objects

To generate all possible hypotheses about the trajectories of the moving objects, we can connect a region  $r_z^b$  to all regions in the superset  $\hat{R}$ ,  $\hat{R} = R_1 \cup \dots \cup R_{z-1}$ , where the edge weight is the similarity score between the connected regions. The optimal trajectories of the moving objects can then be extracted by searching through the entire set of hypotheses and determining the set of optimal paths that best connect these regions. However, as the number of detected regions increases, the search space increases exponentially. In this case, even if we know the starting and ending vertices, it will still be impractical to find these optimal paths. In fact, the tracking problem can be considered as a special case of the NP-complete *Disjoint Connecting Paths* problem. To achieve better computation efficiency while ensuring correct tracking, we propose the following two-phase hypotheses generation and fast hypotheses rejection methods.

#### 4.1. First Phase Hypotheses Generation

In the first phase, we search the set  $R_{z-1}$  for the regions which are similar to a region,  $r_z^b \in R_z$ , in the current frame  $f_z$ . There are two motivations for this scheme. Firstly, if an object is not currently occluded or if its current trajectory

is not split from or merged with the trajectories of other objects, there will not be *sharp* changes in the object's features across successive frames. Consequently, if two regions in successive frames are found to be similar, it is more likely that they are associated with the same object than if the two regions are separated by many frames. Secondly, in the event that there are no region in  $R_{z-1}$  which is similar to  $r_z^b$ , it is likely that there are splits/merges in the objects' trajectories, objects leaving/entering the scene or occlusion of the objects. Since there are many reasons for this ambiguity, we will postpone the decision for the objects' trajectories until more information in later frames is available.

To reduce the false connections, we connect a vertex  $v_{z-1}^a$  to the vertex  $v_z^b$  if they have similar spatial, color and size features. This is expressed in equ. (2),

$$e_{z-1,z}^{a,b} = (\zeta p < \alpha) \cap (\zeta s < \beta) \cap (\zeta h > \gamma) \quad (2)$$

where  $\alpha$ ,  $\beta$  and  $\gamma$  are the thresholds unique to the camera's position, derived by minimizing the Bayesian error.

Despite the criteria for edge connections in equ. (2), a region  $r_z^b$  may be similar to many regions in  $R_{z-1}$ . Hence its corresponding vertex,  $v_z^b$ , may have several parent vertices. To resolve this ambiguity, we refine the criteria for edges connections by equ. (3).

$$e_{z-1,z}^{a,b} = \begin{cases} e_{z-1,z}^{a,b} & \text{if } |F(v_{z-1}^a)| \leq 1 \text{ and } |T(v_z^b)| \leq 1 \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

The derived hypothetical disjoint graph formed after the first phase hypotheses generation is defined as  $G = g_1 \cup \dots \cup g_N$  where  $g_m = (V_m, E_m)$  is a path graph which represents the partial trajectory of an object or a merged group,  $V_m = \{v_i^b, v_{i+1}^c, v_{i+2}^d, \dots, v_{l-1}^x, v_l^y\}$  and  $E_m = \{e_{i,i+1}^{b,c}, e_{i+1,i+2}^{c,d}, \dots, e_{l-1,l}^{x,y}\}$  are the vertex set and edge set of path graph  $g_m$  respectively and  $V_m \cap V_n = \emptyset, m \neq n$ . We denote the set of frames traced by  $g_m$  as  $\Psi(g_m)$ . Our experiments show that by using the methods of edge connections in equs. (2) and (3), we are able to construct these path graphs reliably even when new objects, different from the objects in the training set, enter the scene.

## 4.2. Second Phase Hypotheses Generation

In the second phase, we interconnect the path graphs  $g_1, \dots, g_N$  together to recover the possible trajectories of the objects. We represent the directed weighted edge connecting the end vertex  $v_y^a$  of  $g_j$  to the source vertex  $v_z^b$  of  $g_k$ ,  $j \neq k$ , as  $e_{g_k}^{g_j}$ , where the edge weight is the likelihood that  $g_j$  and  $g_k$  represent two different portions of the trajectory for the same object. We compute the edge weight as follow,

$$w_{g_k}^{g_j} = \frac{\tau}{\zeta p + \zeta s} + \zeta h \quad (4)$$

$$\tau = \frac{\lambda}{z+1-y}, \lambda > 0 \quad (5)$$

where  $\tau$  is use to reduce the influence of the position and size change when the regions represented by  $v_y^a$  and  $v_z^b$  are separated by many frames. This is because the edge  $e_{g_k}^{g_j}$  constructed in the second phase is a hypothesis that  $g_j$  and  $g_k$  represent different portions of the trajectory for the *same* object. If this hypothesis is true, then  $s_y^a$  may be very different from  $s_z^b$  due to the possible large motion of the object during its lost trajectory in frames  $f_{y+1}$  to  $f_{z-1}$ . Similar argument can be made for the position change,  $\zeta p$ . Thus, the influence of  $\zeta p$  and  $\zeta s$  in the calculation of  $w_{g_k}^{g_j}$ , is inversely proportional to the number of frames separating  $r_z^b$  from  $r_y^a$ . On the other hand, since we have normalized the two histograms,  $h_y^a$  and  $h_z^b$ , therefore  $\zeta h$  will not be affected by the number of frames separating them.

Let  $\hat{G}$  be the connected graph formed after the two-phase hypotheses generation. We define the set of path graphs connected to the source vertex and end vertex of  $g_j$  after the two-phase hypotheses generation as  $\Upsilon(g_j)$  and  $\Gamma(g_j)$  respectively, where  $\Upsilon_m(g_j)$  and  $\Gamma_m(g_j)$  denote the  $m^{\text{th}}$  element of  $\Upsilon(g_j)$  and  $\Gamma(g_j)$  respectively and  $w_{g_j}^{\Upsilon_m(g_j)} \geq w_{g_j}^{\Gamma_m(g_j)}, w_{\Gamma_m(g_j)}^{g_j} \geq w_{\Gamma_{m+1}(g_j)}^{g_j}$ .

## 4.3. Fast Hypotheses Rejection Method

The two-phase hypotheses generation method has significantly reduced the search space. Despite this, the search space may still be exponentially large. Hence an intelligent method to perform fast and accurate extraction of the objects' trajectories based on the hypotheses generated is needed to ensure computation efficiency and tracking accuracy. We propose the following fast hypotheses rejection method based on the global information accumulated.

We perform a breath-first transversal of the connected graph  $\hat{G}$  obtained from Sect. 4.2. In the first step, we reject the false hypotheses in  $\Upsilon(g_k)$ . It is most likely that the edge  $e_{g_k}^{g_j}, g_k \in \Gamma(g_j)$  and  $g_j \in \Upsilon(g_k)$ , with the highest weight will be the optimal hypothesis among the hypotheses in set  $\Upsilon(g_k)$ . Consequently, after identifying  $e_{g_k}^{g_j}$ , we can accurately determine the false hypotheses in  $\Upsilon(g_k)$ . Thus, we reasonably assume  $e_{g_k}^{g_j}$  is a valid hypothesis if there are no path graphs  $g_i \in \Upsilon(g_k)$  in which  $w_{g_k}^{g_i} > w_{g_k}^{g_j}$  i.e.

$$e_{g_k}^{g_j} = \begin{cases} 1 & \text{if } \forall_{i \neq j} \nexists g_i, g_i \in \Upsilon(g_k), w_{g_k}^{g_i} > w_{g_k}^{g_j} \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

In the second step, we update the hypotheses in  $\Gamma(g_k)$ . Clearly, the most likely hypothesis in  $\Gamma(g_k)$  is  $g_l$ , where  $w_{g_l}^{g_k} = \max(w_{\Gamma(g_k)}^{g_k})$ . Hence, the next best hypothesis should not have any overlap in its frame number with  $g_l$ . We denote the set of valid hypotheses in  $\Gamma(g_k)$  as  $\Phi^k, \Phi^k \subseteq \Gamma(g_k)$

where,

$$\Phi^k = \Phi_1^k \cup \Phi_2^k \dots \cup \Phi_{|\Gamma(g_k)|}^k \quad (7)$$

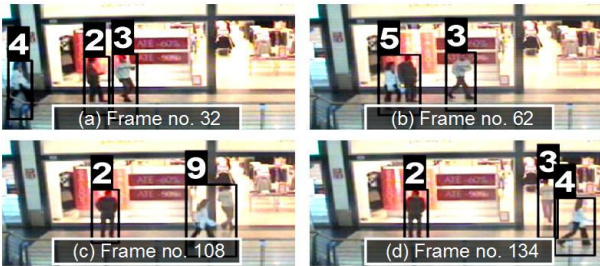
$$\Phi_m^k = \begin{cases} \Gamma_m(g_k) & \text{if } \forall g_i, g_i \in \{\Phi_1^k, \dots, \Phi_{m-1}^k\}, \\ & \Psi(g_i) \cap \Psi(\Gamma_m(g_k)) = \emptyset \\ \emptyset & \text{otherwise} \end{cases} \quad (8)$$

Let  $g_p$  be the path graph that has the closest temporal relations with  $g_k$ ,  $g_p \in \Phi^k$ . In order to achieve continuous tracking of the object,  $g_p$  must be connected to  $g_k$ . We maintain the possible hypotheses of the object by equ. (9).

$$\Gamma(g_p) = \Gamma(g_p) \cup (\Phi^k - g_p) \quad (9)$$

Clearly, our fast hypotheses rejection method achieves the following two goals. Firstly, for each portion of an object's trajectory  $g_k$ , we identify its best hypothesis by relating the hypotheses in the previous frames,  $\{e_{g_k}^{\Gamma_m(g_k)}\}_{m=1}^{|\Gamma(g_k)|}$ , with the hypotheses in the later frames,  $\{e_{\Gamma_m(g_k)}^{g_k}\}_{m=1}^{|\Gamma(g_k)|}$ . This ensures robust tracking. Secondly, the hypotheses rejected in the earlier part of an object's trajectory will not be considered when we reconstruct the later part of the object's trajectory, thus significantly reducing computation complexity. We extract the trajectories of the objects by transversing the path graphs recreated after the fast hypotheses rejection method.

## 5. Experimental Results



**Figure 2.** a) Three persons,  $P_2$ ,  $P_3$  and  $P_4$  enter the scene. b) Detect the merge of the trajectory of  $P_2$  with that of  $P_4$ . c) Trajectory of  $P_2$  extracted following the split of its trajectory from that of  $P_4$ . d) Detect the merge of the trajectory of  $P_3$  with that of  $P_4$ . e) Trajectories of  $P_2$ ,  $P_3$  and  $P_4$  extracted.

We test our multi-object tracking algorithm on five video sequences with a variety of complex motion trajectories such as occlusions and irregular object motions under limited to moderate perspective distortion. There are a total of 46 splits/merges in the trajectories of the objects. We derived the thresholds of equ. (2) by using the first five to fifteen frames of each video sequence. These frames constitute less than 5% of each video sequence.

The only error case is the assignment of a *new* ID to a *previously* detected person following his split from a merged group. 4 such errors were found in the video sequence having moderate perspective distortion of the detected moving regions. Analysis shows that due to perspective distortion, there are sharp changes in the features of the regions in non-successive frames that are associated with the same object. Hence, the region detected following the split from a merged group cannot be matched to any of its previously detected regions. On the other hand, our algorithm achieves 100% tracking accuracy in the video sequences having limited perspective distortion.

One of our tracking results is shown in Fig. 2. We obtain the original test video sequence from [8]. Due to the variations of the lighting condition, pixels in the shadows of the tracked persons are often wrongly detected as foreground pixels. Despite this, we accurately recovered the trajectories of the three individuals after multiple splits and merges. Further demonstrations of our multi-object tracking algorithm are available at <http://perception.i2r.a-star.edu.sg/AlexChia/ReDirect.html>.

## 6. Conclusion and Future Work

In this paper, we present a novel multi-object tracking algorithm based on multiple hypotheses about the trajectories of the objects. Given the detected moving regions of the entire video sequence, we suggest a systematic way to relate these regions by exploiting information in both previous and later frames. Experimental results validate our multi-object tracking algorithm in which we reliably track objects under occlusions, variations of the lighting condition, irregular object motions and multiple splits and merges in video sequences having limited perspective distortion. In the future, we plan to extend this algorithm to track multiple objects over a continuous stream of data.

## References

- [1] I. Cohen and G. Medioni, "Detecting and tracking moving objects for video surveillance," *IEEE Proc. CVPR*, vol. 2, pp. 319–325, 1999.
- [2] H.T. Chen, H. Lin, and T.L. Liu, "Multi-object tracking using dynamical graph matching," *IEEE Proc. CVPR*, vol. 2, pp. 210–217, 2001.
- [3] D.B. Reid, "An algorithm for tracking multiple targets," *IEEE Trans. AC*, vol. 24, no. 6, pp. 843–854, 1979.
- [4] I. Cox and S. Hingorani, "An efficient implementation of reids multiple hypothesis tracking algorithm and its evaluation for the purpose of visual tracking," *IEEE Trans. PAMI*, vol. 18, pp. 138–150, 1996.
- [5] Y.B. Shalom, *Multitarget Multisensor Tracking: Advanced Applications*, Artech House, 1990.
- [6] M. Han, W. Xu, H. Tao, and Y. Gong, "An algorithm for multiple object trajectory tracking," *CVPR*, vol. 1, pp. 864–871, 2004.
- [7] L. Li, W. Huang, Irene Y.H. Gu, and Q. Tian, "Statistical modeling of complex backgrounds for foreground object detection," *IEEE Trans. IP*, vol. 13, no. 11, pp. 1459–1472, 2004.
- [8] EC Funded CAVIAR project/IST 2001 37540, <http://homepages.inf.ed.ac.uk/rbf/CAVIAR/>, Aug 2005.