2010 International Conference on Pattern Recognition

Detection Based Low Frame Rate Human Tracking

Lu Wang, Nelson H.C. Yung
Department of Electrical and Electronic engineering
The University of Hong Kong
Hong Kong, China
{wanglu, nyung}@eee.hku.hk

***Abstract*—Tracking by association of low frame rate detection responses is not trivial, as motion is less continuous and hence ambiguous. The problem becomes more challenging when occlusion occurs. To solve this problem, we firstly propose a robust data association method that explicitly differentiates ambiguous tracklets that are likely to introduce incorrect linking from other tracklets, and deal with them effectively. Secondly, we solve the long-time occlusion problem by detecting inter-track relationship and performing track split and merge according to appearance similarity and occlusion order. Experiment on a challenging human surveillance dataset shows the effectiveness of the proposed method.**

Keywords- low frame rate tracking, data association, ambiguous tracklets, long time occlusion

I. INTRODUCTION

Robust tracking of objects is an important task in video surveillance, which serves as input to high level objects behavior analysis. In many applications, due to the need to reduce the video size for communication and storage, or to be cost efficient, tracking in low frame rate (LFR) is preferred [1]. However, most existing methods cannot be readily applied to LFR tracking because motion is disjoint with a degree of ambiguity. In this paper, we propose a method to associate object detection results into trajectories for LFR tracking.

In data association, researchers tried various means to increase its robustness. [2] proposed a hierarchical framework to progressively resolve the association ambiguity as more information is collected. [3] adapts the weight that multiple cues are combined so as to enhance the discriminative power of the association model in its neighborhood by solving a regression problem. [4] automatically selects features and their corresponding models for association by learning with HybridBoost. In this paper, we propose to improve the linking robustness in LFR tracking by detecting tracklets that introduce either motion or appearance ambiguities.

The framework of generating tracklets first and then linking them by optimization is effective, because the search space for optimization can be reduced significantly without sacrificing much of the accuracy. Typically, 1st-order Markov Chain (MC) assumption is made [2, 4, 5], i.e. the link probability of one tracklet T_i to another tracklet T_j is independent of other tracklets. However, in LFR tracking, this assumption does not hold for a certain amount of tracks, due to the existence of short tracklets, which introduce motion ambiguity because of their lacking of motion prediction probability [Fig 1 (a)]. In addition, when occlusion occurs, the assumption does not hold either, because the appearances of the detections from the occlusion parts are similar to both two related tracks introducing appearance ambiguity, i.e. [Fig. 1 (b)]. The ambiguity can be resolved if three tracklets are considered simultaneously by using 2nd-order MC. However, as 2nd-order MC is very complicated, to efficiently solve this problem, we propose a method that minimizes the use of 2nd-order MC by specifically detecting and dealing with ambiguous tracklets and approximates 2nd-order MC only when it is unavoidable.

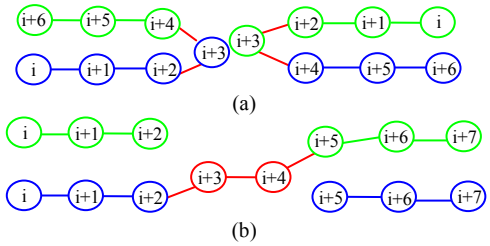


Figure 1. Wrong association caused by ambiguous tracklets. Green and blue circles represent the detection responses of two human objects respectively; red circles are where occlusions occur and the two human objects are mistakenly detected as one; numbers in circles are frame indices; red lines represent incorrect associations.

Long-time occlusion is another issue to deal with. [2] solved the problem by filling the gap between tracklets. However, it fails if the gap is too long, or if only one end of a track is not complete, i.e. there is no gap to fill. [5] solved the problem by performing split-and-merge during tracklets linking. We found that in human surveillance applications, long-time occlusion mostly occurs when human objects are walking together. Therefore, we introduce a track split-and-merge method by measuring inter-track distance in the spatio-temporal space and then merge the tracks, or re-associate the detections to the tracks, according to appearance similarity.

The rest of the paper is organized as follows: in Section II, the proposed LFR data association method is elaborated; Section III talks about how we deal with long-time occlusions; experimental results on a challenging human

1051-4651/10 $26.00 © 2010 IEEE
DOI 10.1109/ICPR.2010.861

surveillance dataset is given in Section IV; and conclusion is drawn in Section V.

II. The Proposed Method

A. Tracklets generation

We first applied our model based method [6] to detect individual human objects and then generate tracklets by linking single responses using the two-threshold strategy proposed in [2]. Link probability is defined as the product of position and appearance affinity. The position affinity is similar to that defined in [7], whereas our appearance affinity of two detection responses r_i and r_j is defined as

$$A_{app}(r_i, r_j) = \frac{\sum_{pt=\{h,t,l\}} \min(VR_{i,pt}, VR_{j,pt}) BC(a_{i,pt}, a_{j,pt})}{\sum_{pt=\{h,t,l\}} \min(VR_{i,pt}, VR_{j,pt})}, \quad (1)$$

where, pt denotes the body part and h, t and l represent head, torso and legs respectively; $VR_{i,pt}$ is the ratio of the number of visible pixels of body part pt of r_i to the whole body size; $a_{i,pt}$ is the 8×8×8 RGB color histogram of part pt of r_i; and $BC(a_i, a_j)$ calculates the Bhattachayya coefficient of the two histograms.

B. Ambiguous tracklets detection

We define two types of ambiguous tracklets. The first type are tracklets having single response. The second type are tracklets having more than one tracklets compete to link to them. Formally, the second type tracklets are defined as: If $p(T_k|T_i)>\theta_1$ and $p(T_k|T_j)>\theta_1$ (both are potential links), $|p(T_k|T_i) - p(T_k|T_j)|<\theta_2$ (but the difference is not significant), T_i and T_j have temporal intersection (i.e. they have at least one detection from the same frame respectively and therefore they cannot be linked to T_k simultaneously), there does not exist any tracklet T_l for T_i (or for T_j) such that $p(T_l \mid T_i)> \theta_1$ (or $p(T_l \mid T_j)> \theta_1$) and T_l and T_k have temporal intersection(if there exists such a tracklet T_l, the ambiguity can be resolved by the Hungarian algorithm), then T_k is defined as an ambiguous tracklet. This is the two-versus-one case and, we can define the one-versus-two case similarly.

C. Tracklets Linking

We propose to link the reliable tracklets first, then, fill missed detections in the resulting tracks with ambiguous tracklets, and lastly perform the complete linking by approximating the 2nd-order MC. All the linking and filling of ambiguous tracklets into tracks are formulated into a MAP framework and solved by the Hungarian algorithm, as is done in [2]. Due to the space limitation, readers are referred to [2] for the formulation details and how Hungarian algorithm is applied to perform data association.

Similar to [2], the link probability p_{link} is defined as the product of three components: appearance, motion and time. Unlike [2], we calculate the appearance of part pt of a tracklet T_i by

$$a_{T_i,pt} = \sum_{k \in T_i} \left(\frac{1}{n-1} \sum_{l \in T_i, l \neq k} A_{app}(r_k, r_l) \right)^2 a_{k,pt}, \quad (2)$$

where n is the tracklet length. The appearance link probability $A_a(T_j|T_i)$ is then calculated by (1) by replacing the apperance and visible ratio of single detections with those of tracks. and the visible ratio of a track is calculated by the right side of (2) by replacing $a_{k,pt}$ with $VR_{k,pt}$.

To calculate motion affinity, because the prediction ability of short tracklets is easy to be affected by inaccurate localization (as shown in Fig. 2), we have to modify it for LFR tracking. For longer tracklets, as Kalman smoothing is applied on the tracklets, the motion affinity becomes more reliable. We define motion affinity $A_m(T_j|T_i)$ as a weighted sum of velocity affinity and orientation affinity as follows.

$$\begin{aligned}
A_m(T_j \mid T_i) &= wA_{m,v}(T_j \mid T_i) + (1-w)A_{m,o}(T_j \mid T_i) \\
A_{m,v}(T_j \mid T_i) &= S_{m,v}(T_j \mid T_i) S_{m,v}(\tilde{T}_i \mid \tilde{T}_j) \\
A_{m,o}(T_j \mid T_i) &= S_{m,o}(T_j \mid T_i) S_{m,o}(\tilde{T}_i \mid \tilde{T}_j)
\end{aligned}$$

$$S_{m,v}(T_j \mid T_i) = \begin{cases} G(\mathbf{p}_i^{tail} + \mathbf{v}_i^{tail}\Delta t; \mathbf{p}_j^{head}, \Sigma_{\Delta t}) & \text{if } |T_i| \geq 3 \\ S_{m,v}(\tilde{T}_i \mid \tilde{T}_j) & \text{if } |T_i| \leq 2 \text{ and } |T_j| \geq 3 \\ \eta^{\Delta t} G(\max(\|\mathbf{p}_j^{head} - \mathbf{p}_i^{tail}\|, v_{max}\Delta t); v_{max}\Delta t, \Sigma_{\Delta t}) & \text{if } |T_i| \leq 2 \text{ and } |T_j| \leq 2 \end{cases}$$

$$S_{m,o}(T_j \mid T_i) = \begin{cases} \sqrt{0.5 + 0.5\mathbf{o}_i^{tail} \cdot \mathbf{o}_j^{head}} & \text{if } |T_i| \geq 2 \text{ and } |T_j| \geq 2 \\ \sqrt{0.5 + 0.5\mathbf{o}_i^{tail} \cdot \frac{\mathbf{p}_j^{head} - \mathbf{p}_i^{tail}}{\|\mathbf{p}_j^{head} - \mathbf{p}_i^{tail}\|}} & \text{if } |T_i| \geq 2 \text{ and } |T_j| = 1 \\ S_{m,o}(\tilde{T}_i \mid \tilde{T}_j) & \text{if } |T_i| = 1 \text{ and } |T_j| \geq 2 \\ 1 & \text{if } |T_i| = 1 \text{ and } |T_j| = 1 \end{cases}, \quad (3)$$

where $A_{m,v}(T_j|T_i)$ is the velocity affinity and $A_{m,o}(T_j|T_i)$ is the orientation affinity; $\tilde{T}_i$ represents the time inversed T_i; $\mathbf{p}_i^{tail}$ (or $\mathbf{p}_i^{head}$) are the filtered real world position of the tail (or head); $\mathbf{v}_i^{tail}$ is the estimated tail velocity of T_i; $\mathbf{o}_i^{tail}$ (or $\mathbf{o}_i^{head}$) are the estimated tail (or head) orientation of T_i, v_{max} is the upper bound of a human object's normal walking velocity; η is a factor penalizing long gap between very short tracklets. In (3), we take tracklets that are shorter than 3 as short tracklets and do not require their velocity prediction accuracy. The orientation prediction ability for tracklets of length two is still used, because generally the orientation, although not accurate, is informative and would not lead to linking errors.

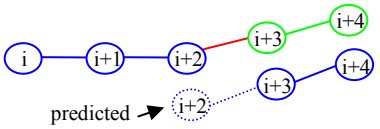


Figure 2. Illustration of unreliable motion prediction of very short tracklets. Due to the inaccurate localization of the response in frame i+3 of the object represeted by blue circles, incorrect linking occur. In such case, motion affinty should be less counted and more emphasis should be put on appearance affinity.

For the temporal affinity calculation, as it aims to differentiate if a gap should be filled or not, we not only exempt the penalty for missed detections caused by

occlusion, but also missed detections where the appearance at the interpolated position is similar to both of the tracklets in consideration.

1) Reliable tracklets linking

When linking the reliable tracklets, we neglect those ambiguous tracklets so as to avoid incorrect linking. This procedure is useful in two folds: first, the linking of the remaining reliable tracklets becomes more robust; second, the speed of calculating the optimal linking is accelerated. The result of this step is illustrated in Fig. 3 (a1) and (b1).

2) Inserting ambiguous tracklets

In this step we fill missed detections of the tracks generated in the last step with ambiguous tracklets. If T_i and T_j are directly linked, but $sf_j - ef_i > 1$ (sf_i and ef_i represent the start frame index and end frame index of T_i respectively), then there are missed detections in this track. The link probability p_{fill} is defined as: for any two tracklets T_i and T_j, $p_{fill}(T_j|T_i)=0$ if T_i and T_j are reliable tracklets but not directly linked; otherwise $p_{fill}(T_j \mid T_i) = p_{link}(T_j \mid T_i)$. It is straightforward to restrict that if T_i and T_j are directly linked reliable tracklets, $p_{term\_fill}(T_i)=0$ (probability of T_i to be an termination end) and $p_{init\_fill}(T_j)=0$ (probability of T_j to be an initialization end).

After obtaining the optimal insertion using the Hungarian algorithm, we might have some incompatible links, i.e. reliable tracklets T_i and T_j are originally directly linked but they are not in the same track after this insertion linking. In this case, we perform a local optimization by breaking the incompatible links and taking T_i as the start tracklet, T_j as the end tracklet, and then using the Hungarian algorithm to link them with the ambiguous tracklets that are not assigned to any other tracks. The result after insertion is illustrated in Fig. 3 (a2) and (b2).

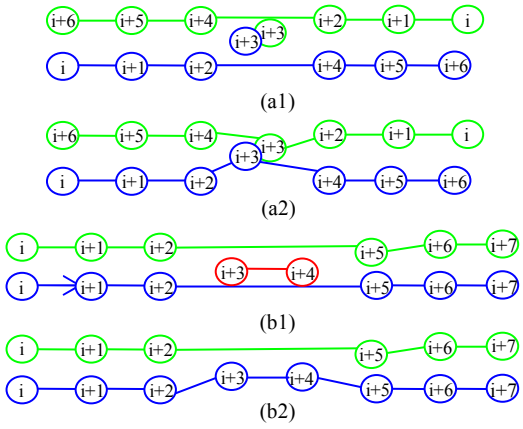


Figure 3. Illustration of the proposed method. (a1) and (b1): reliable tracklets linking; (a2) and (b2): insertion of ambiguous tracklets

3) Approximating 2nd-order MC.

In this step, ambiguous tracklets that are not inserted into tracks in the last step are linked to tracks. Most of these tracklets are extensions to the tracks already formed. As majority of the tracklets have been linked in the previous two steps, the scale of the optimization problem becomes much smaller in this step and we are able to approximate 2nd-order MC without introducing much computational cost. We first do the 1st-order MC linking as is done above, then determine whether there are ambiguous tracklets whose two ends are linked simultaneously: if yes, as this kind of linking may introduce errors, for each such tracklet, break the link at the end with lower link probability; then continue the linking and breaking process until there are no further breaks.

III. DEALING WITH INCOMPLETE TRACKS CAUSED BY LONG-TIME OCCLUSION

This section aims to solve the long-time occlusion. We deal with it after tracklet association by detecting objects moving together. The argument is that, usually, if people are not walking together, occlusion tends to last only a short while. But if people are walking together, once occlusion occurs, its period tends to be long because they move in the same pattern and the occlusion state is unlikely to change within a short time. Define $sf_{i,j}=\max(sf_i, sf_j)$ and $ef_{i,j} = \min(ef_i, ef_j)$. The distance between two tracks T_i and T_j is calculated by

$$D(T_i,T_j)=\frac{\sum_{f=sf_{i,j},\ldots,ef_{i,j}} \left\|\mathbf{p}_{i,f}-\mathbf{p}_{j,f}\right\|}{ef_{i,j}-sf_{i,j}+1}, \quad (4)$$

where $\mathbf{p}_{i,f}$ is the filtered position of detection response of T_i in frame f. If $D(T_i, T_j)$ is smaller than a threshold and the two tracks' coexistent part is long enough, i.e. $ef_{i,j} - sf_{i,j} >= T_f$, human objects represented by the two tracks are considered as walking together.

For each pair of such walk-together tracks, say T_1 and T_2, as illustrated in Fig. 4, we merge them. Suppose T_1 has an extra part than T_2 w.r.t. temporal length. Appearance of each detection in the extra part of T_1 is compared respectively to the single detection appearances of T_1 and T_2 from frames where both tracks have detection responses to support. If the appearance similarity to T_2 is higher than that to T_1 for a certain amount, the detection is assigned to T_2, and thus extends T_2. Otherwise, as no more detection can be assigned to T_2, we decide if T_2 occludes T_1 or T_1 occludes T_2 by comparing the two tracks' distances to the camera. The track with shorter distance to the camera occludes the other one. If T_2 is occluded by T_1, its missing end is merged into T_1; if T_2 occludes T_1, T_2 is left as it is. Further, if a track is merged with other two tracks and those two tracks originally have no temporal intersection, their appearances are compared: if the appearance similarity is high enough, they are identified as the same human object.

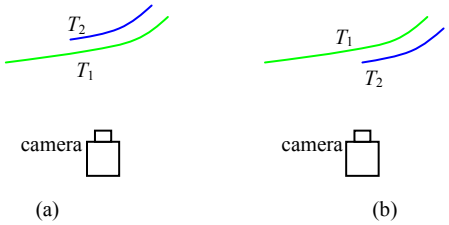
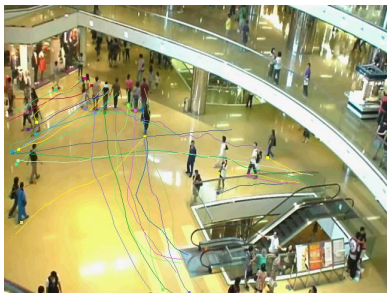


Figure 4. Illustration of walking-together tracks: (a) long track occludes short track; (b) short track occludes long track.

IV. EXPERIMENTAL RESULT

The proposed method is tested on a challenging dataset, which consists of 10 video sequences, with the resolution being 480×640 and the frame rate being 2 frames per second. Each sequence contains 100 frames and there are totally 571 trajectories. Fig. 5 gives an illustration of the scene and the human detection and tracking result. We manually marked our region of interest, which is the central area of the ground, and neglect the detected human objects falling out of the region. The false alarm and missed detection rates of our human detection result are 12.1% and 15.8% respectively.

(a) Human detection result



(b) Associated trajectories

Figure 5. Illustration of the scene and the detection and association result.

TABLE I. EVALUATION OF DATA ASSOCIATE RESULTS

	tracklets	without DAT	with DAT	deal with occlusion
MOTA	0.532	0.747	0.773	0.795

(a1) (a2)

(b1) (b2)

Figure 6. Illustration of association results. (a1) and (a2): result without dealing with ambiguous tracklets; (b1) and (b2): result of the proposed method.

The metric MOTA proposed in [8] is applied to evaluate the association result. Table I lists the scores of the original tracklets, applying the Hungarian algorithm without differentiating ambiguous tracklets (DAT), the proposed method before/after dealing with long-time occlusion.

Fig. 6 demonstrates two associated trajectories generated without and with dealing with ambiguous tracklets. We can see that, although occlusion happens frequently, identity switch is effectively avoided by the proposed method.

In addition, though the proposed method applies the Hungarian algorithm for several times, for the tested data, it is between 4-10 times faster than that without dealing with ambiguous detection. This is because the proposed method can be regarded as a decomposition of the link probability matrix, which effectively reduces the coupling among data.

V. CONCLUSION AND FUTURE WORK

In this paper, we propose to differentiate reliable and ambiguous tracklets when perform LFR data association. Experimental results on a challenging human surveillance dataset shows its effectiveness. The drawback of the method is that if there are too many ambiguous tracklets and too few reliable tracklets, the method might fail because too much information is neglected and other methods have to be developed. Our future work is to use the tracking information to improve the single frame detection accuracy and then refine the tracking accuracy again.

ACKNOWLEDGMENT

This work is supported in part by the Research Grant Council of the Hong Kong Special Administrative Region, China, under Project HKU719608E and in part by the Postgraduate Studentship of the University of Hong Kong.

REFERENCES

[1] F. Porikli and O. Tuzel, "Object tracking in low-frame-rate video," in *SPIE IVCP*, 2005, pp. 72-79.

[2] C. Huang, B. Wu and R. Nevatia, "Robust object tracking by hierarchical association of detection responses," in *ECCV*, 2008, pp. 788-801.

[3] M. Yang, Fengjun Lv, Wei Xu and Yihong Gong, "Detection driven adaptive multi-cue integration for multiple human tracking," in *ICCV*, 2009.

[4] Y. Li, C. Huang and R. Nevatia, "Learning to associate: hybridboosted multi-target tracker for crowded scene " in *CVPR*, 2009, pp. 2953-2960.

[5] A. G. A. Perera, C. Srinivas, A. Hoogs, G. Brooksby and Wensheng Hu, "Multi-object tracking through simultaneous long occlusions and split-merge conditions," in CVPR, 2006, pp. 666-673.

[6] Lu Wang and Nelson Yung, "Crowd counting and segmentation in visual surveillance," in *ICIP*, 2009, pp. 2573-2576.

[7] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by Bayesian combination of edgelet based part detectors," *IJCV,* vol. 75, no 1, pp. 247-266, 2007.

[8] K. Bernardin, A. Elbs and R. Stiefelhagen, "Multiple object tracking performance metrics and evaluation in a smart room environment," in *VS, 2006*.