# SPATIAL-TEMPORAL CONSISTENT LABELING FOR MULTI-CAMERA MULTI-OBJECT SURVEILLANCE SYSTEMS

Jing-Ying Chang*, Tzu-Heng Wang*, Shao-Yi Chien† and Liang-Gee Chen†
*DSP/IC Design Lab., Graduate Institute of Electronics Engineering and Department of Electrical Engineering,
National Taiwan University, Taipei, Taiwan
Email: {jychang, tzwang}@video.ee.ntu.edu.tw
†Email: {sychien, lgchen}@cc.ee.ntu.edu.tw

***Abstract*—For an intelligent multi-camera multi-object surveillance system, object correspondence across time and space is important to many smart visual applications. In this paper, we propose a temporal and spatial consistent labeling algorithm for this demand. In the algorithm, an object corresponding database records the temporal and spatial consistency information for each segmented mask. With the database, the object-mask correlations are propagated through the propagation rules by analyzing mask splitting/merging conditions. In the spatial consistent labeling method, the homography warping and the earth mover's distance are adopted to match same objects across different views. The earth mover's distance solves the double matching problem, allows the algorithm to work normally under a small deviation of detected object locations, and makes pairing results have minimum global matching distances. The concept trusting-former-pairs-more is also adopted to avoid frequent pair switching if two objects are too close. The correct spatial labeling rate is about 89.25% in average. For online processing applications, the algorithm need not trace back to the past frames. The overall processing speed is about 10.24 frame per second (fps) with CIF size video running on a 2.8GHz general purpose CPU.**

## I. INTRODUCTION

With the growing demand for smart environment understanding and market of personal safety, the scale of a surveillance system becomes larger and larger. Indeed, multiple cameras can decrease the dead spots or prevent someone from hiding behind something. We can say we have higher probability to capture every little thing in this system, but the extortionate information also lead to more difficulties in browsing and searching. Besides, before new issues accompanied with multi-camera system like camera installation and data sharing are solved, we can get little benefit from the system.

Intelligent surveillance systems are born to analyze this kind of mass videos and generate compact and meaningful information for human. To build a system like that, one basic but important part is object correspondence. Object correspondence links all projected masks among all camera views constantly if those masks belong to the same object. After building correspondence, not only the routes of objects can be discovered, the systems can further assemble information from every channel to increase the performance of object recognition or behavior analysis.

This paper focuses on deriving object correspondence in a multi-camera multi-object surveillance system. Most of the previous art can be classified into the recognition-based [1] [2] and geometry-based approaches[3][4]. Recognition-based methods adopt the appearance of color or texture patch to do object correspondence. But due to quite different camera natures and quite different viewing angles in every camera, the patches always show different characteristics, which lets a system hard to do object matching between different views with these recognition-based features. Geometry-based approaches use the relationship directly between different views, like disparity, epipolar geometry, and homography. It means if the warping matrix is known, an object in one view can be easily mapped onto the location in the other. That makes geometry-based methods more robust.

This paper is organized as follows. Section II describes the issue and the main concept of our algorithm. In Section II-A and II-A, we present the proposed algorithm, which includes the flows of temporal consistent labeling (TCL) and spatial consistent labeling (SCL). Section III shows subjective and objective results. Section IV is dedicated to concluding remarks.

## II. CONSISTENT LABELING

In a multiple objects and multiple cameras surveillance system, a consistent labeling of an object, which works correctly all the time and cross different camera views, is important for object tracking and knowing interactions between objects. The basic object model extracted from background subtraction and foreground detection is used in the proposed surveillance system. The object model does not record any moving prediction information to prevent prediction errors that occur when two or more objects are too close so as to create a merged mask viewed from a camera. Sometimes, this error will cause tracking failures due to the unpredictable human moving habits. For example, a walking man may turn to another direction when he is entirely occluded by a background object. In this case, the wrong prediction will let the tracker no longer know the position of the man after he reappears from the

978-1-4244-1684-4/08/$25.00 ©2008 IEEE

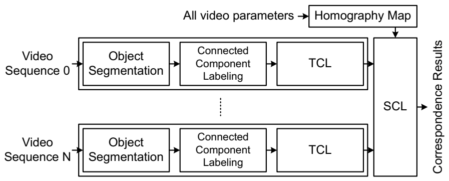


Fig. 1. Spatial-temporal consistent labeling flow.

background object. As indicated in this example, one main issue for object tracking is how to know where objects are during merging, splitting, and reappearing.

The main concept of the proposed consistent labeling algorithm is taking each merged mask as a merged object and recording which single objects are in it. As in Fig. 1, TCL takes charge of the labeling problem throughout consecutive frames for each camera. Then SCL solves the labeling problem across different views with object information from TCL. The two consistent labeling methods are detailed in the following sections.

### A. TEMPORAL CONSISTENT LABELING

The block diagram of TCL is shown in Fig. 2. Every mask in the current frame needs to be compared with the objects saved in a history database. To match a mask with objects in history, four features are extracted to find their similarity. Here we use two different pass thresholds for on-merging masks and other masks. This is because an on-merging mask should be considered as a new merged object. The mask should not be recognized as any single object in the database. These false recognitions happen when one object is too small compared to the other before they are merged, the bigger mask will dominate the on-merging mask and then the similarity checking result. Most of the time, the on-merging mask will match to one of single objects in history if we do not tighten the pass threshold. Hence before object matching stage, the merge/split condition should be detected first to make the matching result favor whether creating a new merged object or selecting an existed object. A merge condition is valid if one mask in the current frame overlaps multiple objects in the previous frame. On the other hand, a split condition is valid if one object in the previous frame overlaps multiple masks in the current frame.

Object matching decision tree is shown in Fig. 3. A mask is matched to one history object with rules based on the following features, color histogram distance, overlapped area size, mask size ratio, and mask centroid distance. Sometimes, there may be multiple matched candidates, and the object with the smallest centroid distance is selected. If no candidates pass all criteria, there would be three kinds of conditions: one is a new object appears; another is the mask is a new on-merging object; the other is the mask appearance changes too fast and fails certain rules. To differentiate which situation happens, the number of objects appearing in the last frame which are overlapped by the current mask indicates the result. If no previous object is overlapped, the mask is identified as a new appearing object. If multiple objects are overlapped, the mask is identified as a new on-merging object. If mask overlaps only one object, it is considered as an object with fast changing appearance.

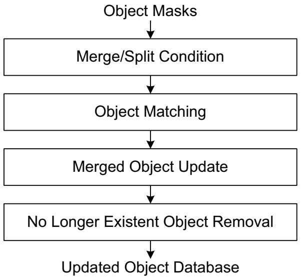


Fig. 2. Temporal consistent labeling block diagram.

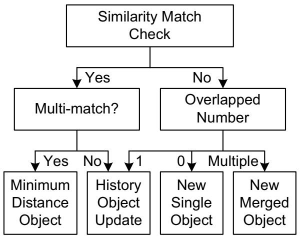


Fig. 3. Object matching decision tree.

After matching all current masks with history objects, some single objects will reappear. If the reappearing objects come from merged objects, the record of which contained objects in the merged objects should be updated. If the number of merged objects of a merged object reduces to zero or one, that means this merged object no longer exists and should be wiped out. Although one mask may be matched to the merged object but not the single object before merge, due to the mask changes a lot during the merging time, this error will be compensated with the step of removing reappearing single objects from merged objects. Finally, to keep a database in a reasonable size, those objects not included in any merged objects and disappearing for a predefined length of time need to be removed from database.

### B. SPATIAL CONSISTENT LABELING

With the assumption that each view should have common ground in most surveillance cases, the proposed algorithm for spatial consistent labeling is based on ground plane homography transformation. Homography transforms the coordinates of a ground point of an object in one view to the coordinates in the other. An example is shown in Fig. 4. Given some matched pairs, we can derive the ground plane warping matrix. The performance of the homography-based consistent labeling significantly depends on the segmentation result and the ground point decision. We still assume the bottom center of a bounding box of an object is its ground point. However, for a mask of a person, the shadow in the mask may shift the true ground point, and for a mask of a vehicle, the different view may have totally different ground points. Two concepts, earth movers distance (EMD) [5] and trusting-former-pairs-more, are employed to prevent from generating wrong matched pairs under this case.

EMD was originally used as a difference measurement between two distributions. In our spatial consistent labeling, the ground point sets in two views are seen as the distributions that need to be compared by EMD. EMD tries to find all matched pairs which will minimize the overall matching distance. That means, if some ground point deviations exist, EMD still works correctly because it finds pairs making global matching distance minimum but not pursuing local (that means one-by-one) matching distance minimum. Besides, EMD has no double matching problem, and we have no need to worry about which pair is closer.

If a new object starts to appear in two cameras, the pair generated at this time has higher confidence than the other pair generated in the future and disobeying this former one. In the beginning that an object appears at a boundary of a view, the object has relatively less occluding problems, less interactions with other objects, and fewer matching choices. At this time, even if it has large ground point deviation, the object still can be paired correctly. But when the object exists in the region of the view longer and longer, all three occurrence frequencies of circumstances increase and then make the generated pair less trustworthy. That makes pairs in history more dependable then new pairs which violates pairs in history. This is the concept of trusting-former-pairs-more. A pair still has a chance to be renewed when the absolute point distance goes too far. In this case, that represents the tracker matches a wrong pair. This situation happens while an object is entering the view but occluded by other objects, which creates an error in TCL and the error propagates to SCL.

The overall flow of SCL is shown in Fig. 5. The database generated by TCL are the input of SCL, and it provides the ground point information of single objects to EMD object matching. The outlier ground points which mean those objects are not shown yet in another view have to be removed before EMD. Those outliers are found and removed according to the absolute distances between their warped points and the ground points of every object in another view. The concept of trusting-former-pairs-more is also realized in this one-to-one object matching block. The merged object masks are not matching in this stage because it might not be a merged object in another view and the ground point of a merged object is meaningless.

Next we find those outliers and disobeying objects in the history pairs to know if we need to create a new pair (paired with nothing) or directly use a pair generated previously. The pairs and coordinates in the database are updated with all information obtained from the previous two stages. Finally, for the merged objects, the contained single objects are used to get the pair information from the database. This stage depends totally on the pairs in the history database.

## III. EXPERIMENTAL RESULTS

The proposed algorithm is tested on PETS2001 dataset 1[6] and two sequences filmed in National Taiwan University. The SCL results are shown in Fig. 6. Color grid points in two views are the given matched points for homography matrix generation. The bounding boxes of objects with same color indicate they are matched pairs. An merged mask will present alternately all color tags come from single objects paired before they are merged. Fig. 7 shows two pair renewal cases mentioned in Section II-B. Fig. 7(a) shows that a man is leaving a car. The system has no prior information about the man in the car and the car is considered as a single object. Due to the warping point deviation, the car in first view matches the man in second view. Later, the distance between them is large enough for fixing their pairs. Fig. 7(b) shows that a car enters first view but it is occluded with the bike at the boundary. That mask is considered as same objects before and after the switching during TCL stage. The error is fixed at the SCL stage.

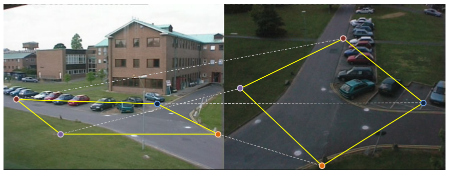


Fig. 4. Warping a ground plane from one view to another using homography transformation.

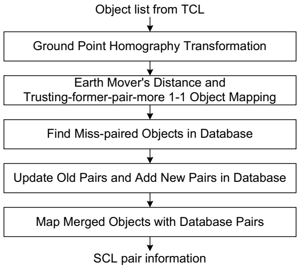

Object list from TCL
Ground Point Homography Transformation
Earth Mover's Distance and
Trusting-former-pair-more 1-1 Object Mapping
Find Miss-paired Objects in Database
Update Old Pairs and Add New Pairs in Database
Map Merged Objects with Database Pairs
SCL pair information

Fig. 5. Spatial consistent labeling block diagram.

The objective experimental results are shown in Table I. The author, Kevin Smith [7], defines four evaluating metrics for temporal consistent labeling. They are average object purity($\overline{OP}$), average tracker purity ($\overline{TP}$), average false identified object ($\overline{FIO}$), and average false identified object ($\overline{FIT}$). $\overline{OP}$ stands for how much percentage that ground true objects are detected by trackers. $\overline{TP}$ stands for how much percentage that trackers detect ground true objects correctly. $\overline{FIO}$ stands for average number of falsely identified object per frame. $\overline{FIT}$ stands for average number of falsely identified tracker per frame. Because all $\overline{FIT}$ errors in our experiments come from the wrong tracker switching, $\overline{FIT}$ and $\overline{FIO}$ will have the same values. The concept of these metrics is then modified for SCL. We derive the ratio of the number of false pairs to the number of total pairs over entire sequence and call it average false identified pair ($\overline{FIP}$). The average $\overline{FIP}$ of these three sequences is $0.1075$, which means there are $0.1075$ false pairs per one ground true pair. That means the correct spatial labeling rate is about $89.25\%$ in average. Most errors

TABLE I
SPATIAL AND TEMPORAL CONSISTENT LABELING RESULTS

| Test Sequences | View | $\overline{\text{TP}}$ | $\overline{\text{OP}}$ | $\overline{\text{FIT}}$($\overline{\text{FIO}}$) | $\overline{\text{FIP}}$ |
|---|---|---|---|---|---|
| PETS2001 (Outdoor) | 1 | 99.32% | 94.58% | 0.0057 | |
| | 2 | 98.34% | 81.25% | 0.0500 | |
| | Avg. | 98.81% | 87.23% | 0.0279 | 0.2168 |
| Ming-Da Hall (Indoor) | 1 | 77.43% | 78.94% | 0.1810 | |
| | 2 | 87.37% | 84.34% | 0.0076 | |
| | Avg. | 82.40% | 81.64% | 0.0943 | 0.0294 |
| Barry Lam Hall (Outdoor) | 1 | 100.00% | 90.55% | 0.0000 | |
| | 2 | 100.00% | 88.57% | 0.1199 | |
| | 3 | 100.00% | 86.47% | 0.0000 | |
| | 4 | 100.00% | 96.74% | 0.0000 | |
| | Avg. | 100.00% | 90.58% | 0.0300 | 0.0763 |

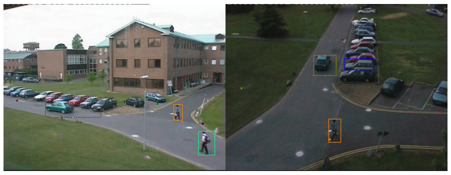
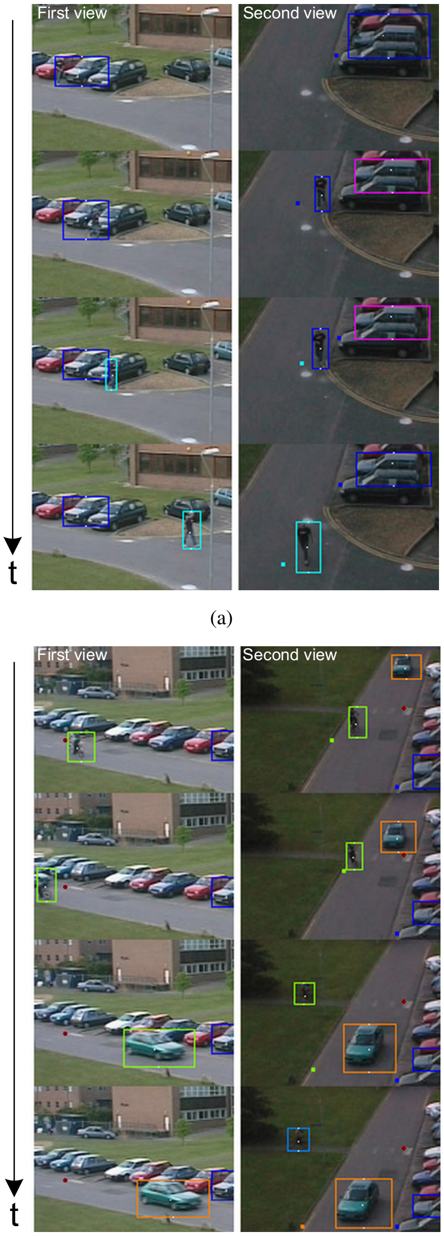


Fig. 6. Spatial consistent labeling results. The green car contains green and red tags because the driver goes into the car previously. The merged mask in the first(left) view contains green, red, and blue tag which represents the driver, the green car, and the dark blue car in turn. If an object in the first view is a single object, its transformed ground point with the same color tag is shown in right view.

are generated from bad foreground masks and the situation when objects enter the view but occluded by others. Running with Intel Core 2 Duo processor at 2.8 GHz, the SCL speed is about 4,638 fps per one pair of CIF size video channels. The overall speed is about 10.24 fps.

## IV. CONCLUSION

An effective TCL and SCL algorithm is proposed in this paper. The objects within an occluded mask need not to be separated to make our TCL and SCL work fine. Homography ground plane warping is used for deriving SCL. EMD is utilized to do object matching, and the concept of trusting-former-pairs-more is applied to generate the matching pairs across temporal and spatial domain correctly and stably. $\overline{FIP}$ metric is also defined for an objective judgement on the performance of the proposed SCL algorithm. As indicated in the experimental results, the $\overline{FIP}$ is 0.1075 which means about 89.25% of pairs are correctly identified. For online processing applications, the proposed algorithm need not trace back to the past frames.



(a)

(b)

Fig. 7. Two pair renew examples. Fig. 7(a) shows that a man is leaving a car. Fig. 7(b) shows that a car enters left view.

## REFERENCES

[1] J. Krumm, S. Harris, B. Meyers, B. Brumitt, M. Hale, and S. Shafer, "Multi-camera multi-person tracking for easyliving," in *IEEE International Workshop on Visual Surveillance*, 2000, pp. 3–10.

[2] A. Mittal and L. Davis, "M2tracker: A multi-view approach to segmenting and tracking people in a cluttered scene using region-based stereo," in *European Conference on Computer Vision*, 2002, pp. 18–36.

[3] W. Hu, M. Hu, X. Zhou, T. Tan, J. Lou, and S. Maybank, "Principal axis-based correspondence between multiple cameras for people tracking," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 28, no. 4, pp. 663–671, 2006.

[4] S. Khan and M. Shah, "Consistent labeling of tracked objects in multiple cameras with overlapping fields of view," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 25, pp. 1355–1360, 2003.

[5] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," Stanford, CA, USA, Tech. Rep., 1998.

[6] "Pets2001 dataset," http://www.cvg.cs.rdg.ac.uk/PETS2001/pets2001-dataset.html.

[7] K. Smith, D. Gatica-Perez, J. Odobez, and B. Sileye, "Evaluating multi-object tracking," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 3, 2005, pp. 36–43.