21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

Semantic Superpixel based Vehicle Tracking

Liwei Liu[1], Junliang Xing[1], Haizhou Ai[1] and Shihong Lao[2]
[1]*Computer Science and Technology Department, Tsinghua University, Beijing, 100084, China*
[2]*Development Center, OMRON Social Solutions Co., LTD, Kyoto 619-0283, Japan*
E-mail:ahz@mail.tsinghua.edu.cn

Abstract

This paper focuses on tracking multiple vehicles in real-world traffic videos which is very challenging due to frequent interactions and occlusions between different vehicles. To address these problems, we fall back on superpixel which recently has received great attention in a wide range of vision problems, e.g. object segmentation, tracking and recognition, for its ability of capturing local appearance characteristics of objects and their spatial relations. As a mid-level feature, however, superpixel itself is unable to carry semantic information which may restricts their use in these problems. To this end, we introduce semantic information into superpixel from an offline trained semantic object detector and successfully deploy it into the multiple vehicle tracking problem. The benefits of semantic superpixel include: 1) it gains better temporal coherency of superpixel; 2) the effectiveness and robustness of occlusion handling are improved; 3) benefited from semantic analysis, false targets and false trajectories are significantly reduced. Experiments show significant accuracy improvements of our approach in comparison with existing tracking methods.

1. Introduction

Multiple objects tracking in video is of fundamental importance for surveillance system and provides great potentials for many applications, such as visual surveillance, traffic safety monitoring, intelligent scheduling, etc. The difficulties behind multiple object tracking, however, are also pronounced, the major one of which is occlusions between different objects which are often encountered and cause tracking failure.

The performance of tracking algorithm depends much on observations of targets. According to different observations, the state-of-the-art approaches can be mainly categorized into three classes: low-level, mid-level and high-level observation based approaches. Although low-level observations have great potentials in occlusion handling, they are usually unstable. Kanade-Lucas-Tomasi (KLT) [9] tracks corner features frame-to-frame, however, it is unreliable under conditions of video blur or abrupt motion. ST-MRF based block propagation [6] requires vehicles to be identified separately before occlusion happens. Similar adjacent pixels are clustered into superpixels to compute local features. As mid-level observation, superpixel has been proven to be effective representation in image segmentation [4] and object recognition [8]. In [11], a discriminative appearance model based on superpixels is presented to facilitate a tracker to distinguish the target and the background. But it may lose effectiveness when background is similar or multiple targets appear, since it neglects of semantic information. Recently, the fast development of object detection techniques have resulted in many promising detectors, e.g., faces [10][5], pedestrians [3][12] and vehicles [7]. These object detectors provide good observation models for detection based tracking algorithm. Although these high-level observation based approaches facilitate themselves with global description of target classes, they are very sensitive to object occlusions.

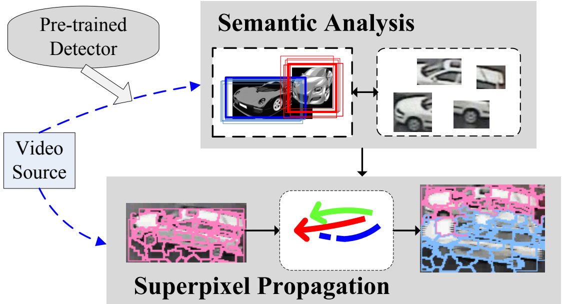

Figure 1: System overview.

In this paper, we focus on multiple vehicles tracking with interactional occlusions which is quite common in surveillance application. As shown in Fig. 1, superpixel propagation is utilized to exploit effective visual cues for vehicle tracking with occlusions. Semantic analysis guides the superpixels to model and track targets in semantic level (semantic superpixel), which performs superpixel propagation smoothly and accurately to maintain temporal coherency, and furthermore reduce false targets and false trajectories greatly.

2. Semantic Superpixel Tracker

Semantic Superpixel Tracker (SST) couples semantic analysis and superpixel propagation tightly together to multi-target tracking. Although our superpixel propagation is effective of handling occlusions, it is ignorant of what an object is, which may lead to the failure of multiple targets tracking. In this section we first reveal the deficiency of superpixel propagation in case of multi-target tracking, and then introduce the details of semantic analysis, finally describe how the semantic analysis guides the superpixel propagation.

2.1. Semantic Analysis

Superpixel propagation is unable to ensure that the group of superpixels it tracked is a part of a target or adjacent targets or a true target. When the group of superpixel becomes disconnected, there may be two possibilities: 1. the disconnected parts are different true targets since the targets entered the scene adjacently and were mistaken as a target; 2. the disconnected parts belong to a single target due to occlusions or bad foreground extraction. Since low-level cues cannot catch which case it is, without semantic guidance, it will eventually leads to tracking loss. Therefore, semantic analysis is employed to guide the superpixel propagation, in which detectors are adopted to exploit semantic information. The semantic consists of two aspects: reconstruct superpixel groups and cooperate with vehicle components to provide accurate motion vector.

Superpixel group reconstruction reconstructs superpixel groups based on detection responses, which guarantees that each group is a true target. We propose a sampling process to incorporate the semantic information into superpixel grouping. As semantic information provider, our detector will output a value for an image region. The value, called confidence, describes how similar to a true target is the image region (the higher confidence, the more likely it corresponds to a true target). Based on this, we first gaussianly sample n particles (image regions) in each superpixel group to capture the responses of true targets. And then merge the particles of high confidences into clusters C_i. Therefore the state of the superpixel group $\mathbf{X_t}$ will be divided into new states $\mathbf{X}_{t,C_i}$ according to the clusters C_i:

$$\mathbf{X}_{t,C_i} = \sum_{\mathbf{x}_t^n \in C_i} \omega^n \mathbf{x}_t^n \tag{1}$$

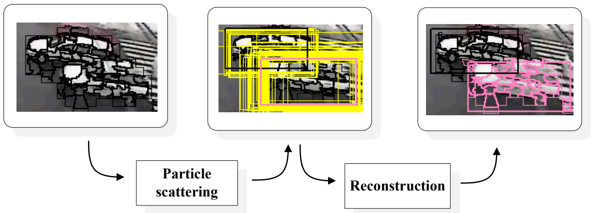

Figure 2: Superpixel group reconstruction.

where $\mathbf{X}_{t,C_i}$ is the new state associated with C_i and $\mathbf{x}_t^n$ is a set of weighted ω^n particles that belong to C_i. Superpixels will be redistributed according to $\mathbf{X}_{t,C_i}$. Two aspects should be noted here: 1) when two targets overlap, the overlap sub-regions will be redistribute to the one with higher visible score which is the mean particle confidence of the target; and 2) To prevent detectors from drifting when vehicles are close to each other, we remove the new state $\mathbf{X}_{t,C_i}$ conflicted with existing targets whose states are predicted from previous states and motion models. The weight ω^n of particle $\mathbf{x}_t^n$ can be computed as following:

$$\omega^n = \frac{\pi^n}{\sum_{\mathbf{x}_t^j \in C_i} \pi^j} \tag{2}$$

where π^n is the redefined confidence for particle $\mathbf{x}_t^n$ since detector outputs are inaccurate. Considering the characteristic of a boosted object detector, the redefined confidence is calculated as:

$$\pi = \frac{\eta \exp(S)}{\exp\left(L_T - L_P\right)} \tag{3}$$

where L_T is the total layer number of the cascade detector, L_P is the layer number the particle passed, S is the sum of marginal distances (differences between detector outputs and offline learnt thresholds in each layer) when passing one layer, and η is a normalization factor.

During the tracking procedure, detectors acted as semantic information provider also play a crucial role in our framework. **Split**: when a group of superpixels (with the same target ID) splits into disconnected parts, should we assign the disconnected regions with new ID and track them respectively or maintain the old ID and track them as a single target? In the same token, the higher redefined confidence sum of scattered particles is, the more likely it is a true target. Therefore, the scheme with higher redefined confidence sum will be adopted. **Merge**: on rare occasions, a group of superpixels which is a part of a true target is tracked as another target. To handle with this, we merge these superpixels into the target which contains them for 5 frames.

Motion vector calculation is carried on with the help of mid-level cues and detection results. The performance of superpixel propagation depends on the accuracy of motion model to a large extent. With accurate predicted motion, superpixels will maintain their characteristic over time. But when a region is similar with its neighboring regions, the motion calculated based on SSD will be unreliable. So before motion calculation, it is necessary to cluster superpixels into optimal regions (vehicle components) which have large discriminative ability from their neighboring regions.

Given neighboring superpixels S_i and S_j, we define their dissimilarity $D(S_i, S_j)$. To calculate the dissimilarity, multiple cues, such as color, edge and texture information, can be fused together to measure the metric between the two superpixels. In our experiment, we design the following measurement that is effective for our problem and yet of high computation efficiency:

$$\begin{aligned} D\left(S_i, S_j\right) &= \left|\operatorname{Mean}\left(S_i\right) - \operatorname{Mean}\left(S_j\right)\right| \\ &+ \left(\operatorname{Var}\left(S_i \cup S_j\right) - \operatorname{Var}\left(S_i\right) - \operatorname{Var}\left(S_j\right)\right) \\ &+ G\left(S_i \cap S_j\right) \end{aligned} \tag{4}$$

The first term calculates the color distance between the mean colors of the superpixels, the second term calculates the variance changes after merging S_i and S_j, and the third term calculates the average gradient magnitude of the boundary. With the dissimilarity measurements, we progressively merge neighboring superpixels that have the minimal dissimilarity values,

- Sort the D values;
- Pick up neighboring superpixels with the lowest D values and merge them;
- In order to prevent superpixels over-merging, the procedure is carried on until the lowest D value exceeds a threshold.

After superpixel merging procedure, we are left with discriminative regions and SSD is applied to calculate motion vectors (search for the most similar region in next frame using dynamic programming). As for a vehicle (rigid object), all pixels within it have the similar motion vectors. Therefore, we get the superpixels within each detection bounding box and calculate their mean motion vector as the motion vector for both the target and the superpixels.

2.2. Semantic Superpixel Propagation

In order to track superpixels in image sequences, it is crucial to exploit prior knowledge obtained during the processing of the previous images. We must guarantee that tracked superpixels have similar characteristics with previous superpixels. Therefore, we adopt superpixel propagation to track superpixels, in which the segmentation result from time $t-1$ will be the initialization of the segmentation to be computed at time t:

$$S_{t\_}^i = \left\{x_{t-1}^i + \Delta x, y_{t-1}^i + \Delta y, L_{t-1}^i, a_{t-1}^i, b_{t-1}^i\right\} \tag{5}$$

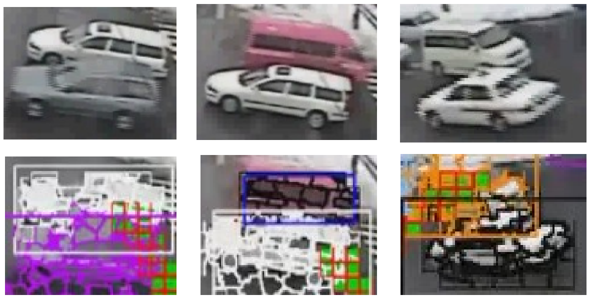

Figure 3: Sample results of semantic superpixel propagation when detectors lose effectiveness (1st row: source images; 2nd row: the propagation results).

$S_{t\_}^i$, the ith superpixel initialization at time t, contains five dimension data: its center coordinate at previous frame $\left(x_{t-1}^i, y_{t-1}^i\right)$ and the motion vector $(\Delta x, \Delta y)$ from last section provide the initial center for $S_{t\_}^i$; $\left\{L_{t-1}^i, a_{t-1}^i, b_{t-1}^i\right\}$ is the initial color values from previous frame in Lab color space. We employ SLIC [1] to implement superpixel segmentation. In short, our superpixel propagation move superpixels from previous frame to current frame, and then use their information as initialization to carry out superpixel segmentation. With initialization, the superpixel segmentation significantly decreases the computational effort. Furthermore the superpixels will almost be coherent over time.

After superpixel propagation, S_t^i will inherit the target id from S_{t-1}^i. Target O_{id} updates its state from $O_{t-1,id}$ to $O_{t,id}$ based on its superpixel propagation:

$$\left\{S_{t-1,id}^1, S_{t-1,id}^2, ..., S_{t-1,id}^n\right\} \rightarrow \left\{S_{t,id}^l, ..., S_{t,id}^m\right\} \tag{6}$$

When occlusion happens, visible superpixels will propagate successfully and occluded ones will be lost ($m \leq n$). Therefore, superpixel propagation has great power in capturing visible cues to handle with occlusions, just as shown in Fig. 3. When occluded parts reappear gradually, new superpixels should be generated to capture these parts. We employ consecutive frames subtraction in foreground to obtain these parts, and then seeds will be scattered in these parts, after that SLIC superpixel segmentation is carried out to achieve the new superpixels. According to neighborhood relationship and the semantic analysis, we finally assign target id to them.

3. Experiments

Experiments are carried out on the evaluation datasets used in [7] and three challenging real-world videos collected with a hand-held camera. The sequences contain frequent interactions and occlusions between vehicles.

Table 1: Tracking Comparison

Method	GT	MT	ML	FRMT	FAT	IDS
Ours	215	198	4	17	7	3
Liu et al. [7]	215	187	6	41	3	5
ST-MRF [6]	215	193	4	22	31	6

3.1. Experiment Settings

The vehicle detectors are offline trained in the boosting framework with Joint Sparse Granular Features (JSGF) [2] which has been proven to be effective for vehicle detection [7]. We obtain foreground by background subtraction [6] and do superpixel segmentation and propagation in foreground. The algorithm also sets up a start line at each entrance. When a connected foreground region free of ID goes through the start lines, it will be assigned a new ID, and seeds will be scattered equably in foreground region to process initialized superpixel segmentation [1].

3.2. Tracking Performance Evaluation

We adopt the same metrics for evaluating tracking performance as in [12]. These metrics are defined as following. MT: number of Mostly Tracked trajectories; ML: number of Mostly Lost trajectories; FRMT: number of Fragments trajectories; FAT: number of False trajectories; IDS: the frequency of Identity Switches.

We compare our approach with the method in [7] and ST-MRF in [6] (implemented by ourselves). Table 1 gives the comparison results. Liu et al. [7] employs high-level observations to track vehicles which cannot handle occlusions well due to its congenital deficiency at seizing local features. From the table, we can see that our approach achieve significant improvements on FRMT and MT since our approach is robust in occlusion handling and reduce the number of fragments trajectories greatly. On the other extreme, [6] focuses on low-level cues but neglects of semantics which lead to many false trajectories (FAT). With semantic feedback and corresponding operations, our approach can reduce FAT to a great extent. We attribute these significant improvements to the effective fusion of semantic analysis and superpixel propagation of our approach. Some typical results are shown in Fig. 4.

4. Conclusion

In this paper, we propose a robust multiple occluded vehicle tracking algorithm in common traffic surveillance environments. Mid-level and high-level observations are tightly coupled together to track vehicles with occlusions. With semantic guidance derived from detectors, our superpixel propagation not only improves its power in occluded vehicle tracking but also overcomes the drawback of high false alarms. Experiment results on traffic surveillance datasets and real-world video data demonstrate the effectiveness of our approach.

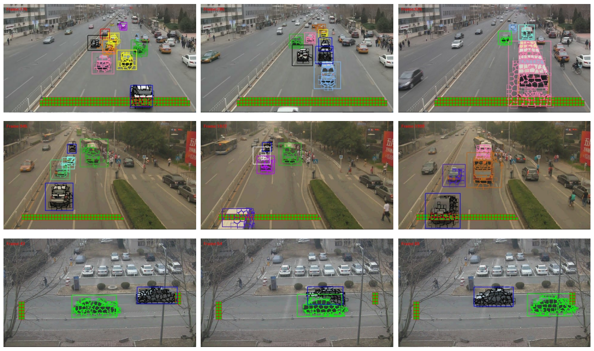

Figure 4: Some typical results with occlusions.

References

[1] R. Achanta, A. Shaji, K. Smith, and A. Lucch. Slic superpixels. Technical Report 49300, EPFL, 2010.

[2] H. Ai, C. Huang, S. Lao, and T. Yamashita. Specified object detection apparatus, 2007.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[4] B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *ICCV*, 2009.

[5] C. Huang, H. Ai, Y. Li, and S. Lao. High performance rotation invariant multiview face detection. *PAMI*, 29:671–686, 2007.

[6] S. Kamijo, Y. Matsushita, K. Ikeuchi, and M. Sakauchi. Occlusion robust tracking utilizing spatio-temporal markov random field model. In *ICPR*, 2000.

[7] L. Liu, J. Xing, and H. Ai. Multi-view vehicle detection and tracking in crossroads. In *ACPR*, 2011.

[8] C. Pantofaru, C. Schmid, and M. Hebert. Object recognition by integrating multiple image segmentations. In *ECCV*, 2008.

[9] C. Tomasi and T. Kanade. Detection and tracking of point features. Technical Report CMU-CS-91-132, Carnegie Mellon University, 1991.

[10] P. Viola and M. J. Jones. Robust real-time face detection. *IJCV*, 57:137–154, 2004.

[11] S. Wang, H. Lu, F. Yang, and M. Yang. Superpixel tracking. In *ICCV*, 2011.

[12] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *IJCV*, 75:247–266, 2007.