Online Multi-Object Tracking and Segmentation with GMPHD Filter and Simple Affinity Fusion

Young-min Song and Moongu Jeon
Gwangju Institute of Science and Technology
Gwangju 61005, South Korea
{sym,mgjeon}@gist.ac.kr

Abstract

In this paper, we propose a highly practical fully online multi-object tracking and segmentation (MOTS) method that uses instance segmentation results as an input in video. The proposed method exploits the Gaussian mixture probability hypothesis density (GMPHD) filter for online approach which is extended with a hierarchical data association (HDA) and a simple affinity fusion (SAF) model. HDA consists of segment-to-track and track-to-track associations. To build the SAF model, an affinity is computed by using the GMPHD filter that is represented by the Gaussian mixture models with position and motion mean vectors, and another affinity for appearance is computed by using the responses from single object tracker such as the kernalized correlation filters. These two affinities are simply fused by using a score-level fusion method such as Min-max normalization. In addition, to reduce false positive segments, we adopt Mask IoU based merging. In experiments, those key modules, i.e., HDA, SAF, and Mask merging show incremental improvements. For instance, ID-switch decreases by half compared to baseline method. In conclusion, our tracker achieves state-of-the-art level MOTS performance.

1. Introduction

Multi-object tracking (MOT) has been an emerging research field in the last decade while the representative MOT benchmark datasets [1, 4, 9] have been released and simultaneously tracking-by-detection paradigm has been exploited as top trend for MOT. Also, breakthroughs in object detection have been achieved by many deep neural networks (DNN) based detectors [10, 12, 13, 14] from various sensor domains such as color camera (2D images) and LiDAR (3D point clouds), respectively. According to those input sources, the detectors give different outputs, i.e., observations. For instance, detection responses of [12, 13] are 2D bounding boxes and those of [10, 14] are 3D boxes. In addition, K. He *et al*. [5] introduced a pixel-wise classification and detection method, represented by instance segmentation, which has motivated many segmentation based researches. Figure 1 shows examples of those three kinds of detections results.

Figure 1. Examples of detection results in KITTI dataset which are visualized for the same image. (a) and (c) were obtained from Regionlets [19] and Mask R-CNN [5] with the camera input, respectively. (b) was obtained from Point R-CNN [14] with LiDAR 3D point clouds input and calibrated to the camera image coordinates. Each object has a class number indicating car or pedestrian.

Accordingly, a new MOT task has been most recently explored aiming for pixel-wise intelligent systems beyond 2D bounding boxes which is named multi-object tracking and segmentation (MOTS) that was first introduced in Voigtlaender *et al*. [18] with new evaluation measures and a new baseline method. They also released a new dataset extended from KITTI [4] and MOTChallenge [9] image sequences. Luiten *et al*. [11] proposes a MOTS method which uses fusing of 2D box detection, 3D box detection, and instance segmentation results. Motivated from these MOTS works and other conventional MOT researches, we propose a highly

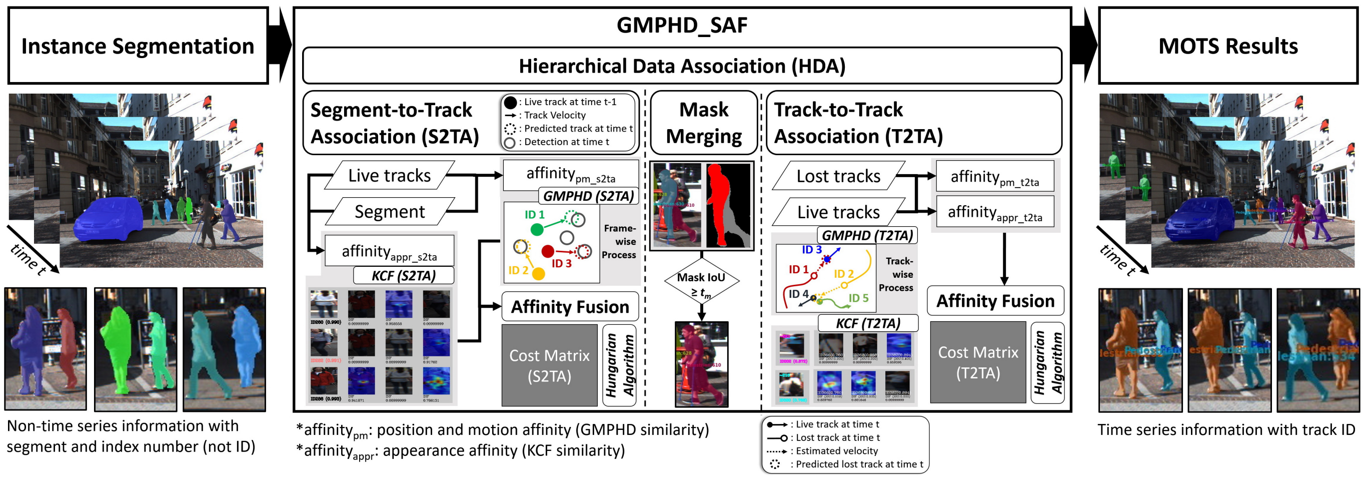

Figure 2. Processing pipeline of GMPHD_SAF with input (images and instance segmentation results) and output (MOTS results). Key components are Hierarchical data association (HDA), Mask merging, and Simple affinity fusion (SAF). HDA has two association steps: S2TA and T2TA. SAF executes each affinity fusion in each association step while Mask merging runs once between S2TA and T2TA.

practical online MOTS method in this paper. Our contributions are summarized as follows:

1) We propose a highly practical online MOTS method which is based on (a) the GMPHD filter and consists of (b) Hierarchical data association (HDA), (c) Mask merging, and (d) Simple affinity fusion (SAF). These four modules successfully build a feasible online MOTS framework.

2) We evaluate the proposed method on a state-of-the-art datasets [18]. Evaluation results on the training sets show incremental improvements compared to a baseline method. In the results on test sets, our method not only shows the best performance against state-of-the-art published methods but also achieves state-of-the-art level performance against state-of-the-art unpublished methods which are available at the leaderboards of KITTI-MOTS and MOTSChallenge websites.

We introduce the proposed method in Section 2 in detail and discuss the experimental results in Section 3, and conclude this paper in Section 4. From now on we will use GMPHD_SAF as the abbreviation for the proposed method.

2. Proposed method

The GMPHD filter [17] has been widely used for online approach in state-of-the-art 2D box MOT methods [2, 3, 8, 16]. Thus, we exploit it for online multi-segment tracking i.e., MOTS. GMPHD_SAF consists of four key components: the GMPHD filter based tracking process, hierarchical data association (HDA), Mask merging, and Simple affinity fusion (SAF). In this section, we address what inputs/outputs those key modules work with in HDA, how position and motion affinity and appearance affinity are fused by SAF, and what metric is used for Mask merging, as described in Figure 2.

2.1. The GMPHD filter

The main steps of the GMPHD filtering based tracking includes Initialization, Prediction, and Update which are introduced in supplementary material [15] in detail.

Observations (instance segmentation) and states (segments tracks) at time t are represented as follows:

$$X_t = \{\mathbf{x}_t^1, \ldots, \mathbf{x}_t^{N_t}\}, \tag{1}$$

$$Z_t = \{\mathbf{z}_t^1, \ldots, \mathbf{z}_t^{M_t}\}, \tag{2}$$

where a state vector $\mathbf{x_t}$ is composed of $\{x, y, vx, vy\}$ with track ID, and segment mask. x, y, and vx, vy indicate the center coordinates of the mask's 2D box, and the velocities of x and y directions of the object, respectively. An observation vector $\mathbf{z_t}$ is composed of $\{x, y\}$ with segment mask. A Gaussian model representing $\mathbf{x_t}$ is initialized by $\mathbf{z_t}$, predicted to $\mathbf{x_{t+1|t}}$, and updated into $\mathbf{x_{t+1}}$ by $\mathbf{z_{t+1}}$.

2.2. Hierarchical data association

HDA has two-step association: Segment-to-track association (S2TA) and Track-to-track association (T2TA). Each association has different observations and states as inputs to compute affinity$_{pm}$ and affinity$_{appr}$, see Figure 2.

Segment-to-track association. Inputs at time t are equal to observations (2) and states (1).

Track-to-track associations. Observations and states (inputs) are live tracks and lost tracks. Live and lost tracks' vectors have the information of (1) with birth time t_b and

Trackers		Modules				KITTI-MOTS Training Sequences							
		S2TA	Mask Merging		T2TA	Cars				Pedestrians			
		SAF	IoU	Mask IoU	SAF	sMOTSA↑	MOTSA↑	IDS↓	FM↓	sMOTSA↑	MOTSA↑	IDS↓	FM↓
Ours	$p1$					73.7	84.0	1322	1250	56.4	71.2	800	721
	$p2$	✓				76.3	86.6	642	606	59.6	74.5	428	387
	$p3$	✓	✓			76.8	86.5	598	572	59.5	74.3	429	391
	$p4$	✓		✓		77.0	86.7	581	557	59.6	74.4	423	382
	$\boldsymbol{p5}$	✓		✓	✓	**77.8**	**87.6**	**362**	**518**	**61.2**	**76.0**	**245**	**341**

Table 1. **Evaluation results on KITTI-MOTS training sequences**. In $p3$, $p4$, and $p5$, merging threshold t_m is set to 0.4.

Trackers		MOTSChallenge Training Sequences			
		Pedestrians			
		sMOTSA↑	MOTSA↑	IDS↓	FM↓
Ours	$p1$	64.5	75.9	686	604
	$p2$	64.5	75.9	535	**487**
	$p3$	64.6	75.9	565	523
	$p4$	65.0	76.3	539	497
	$\boldsymbol{p5}$	**65.6**	**77.1**	**335**	509

Table 2. **Evaluation results on MOTSChallenge training set**. In $p3$, $p4$, and $p5$, merging threshold t_m is set to 0.4.

lost time t_l. Live track's t_b is identical to current time t and t_l is not assigned yet. Lost track's t_l is less than t that means the track is lost before the current time.

By using these inputs in S2TA and T2TA, affinity (cost) matrices are computed and we use the Hungarian method [7] to solve the cost matrices. Then, some observations are assigned to associated states for update, and other non-assigned observations initialize new states.

2.3. Simple affinity fusion

Fusing affinities obtained from different domains requires a normalization step which can balance the different affinities and avoid bias by one affinity which may have higher magnitude than others.

Position and motion affinity. In fact, the GMPHD filter includes Kalman filtering that designs prediction by using linear motion with noise. Therefore, position and motion affinity between i_{th} state and j_{th} observation gives the probabilistic value $w \cdot q(\mathbf{z})$ by the GMPHD filter as follows:

$$A_{pm}^{(i,j)} = w^i \cdot q^i(\mathbf{z}^j), \tag{3}$$

which is acquired in Update step of the GMPHD filter [15].

Appearance affinity. We exploit the Kernelized correlation filter (KCF) [6] for computing appearance affinity between i_{th} state and j_{th} observation. The affinity can be derived as follows:

$$A_{appr}^{(i,j)} = 1 - \frac{\sum_{c=x_j}^{width_j} \sum_{r=y_j}^{height_j} \bar{d}_{KCF}^{(i,j)}(r,c)}{width_j \cdot height_j}, \tag{4}$$

where $\bar{d}(\cdot)$ indicates the normalized KCF distance value that has ranges 0.0 to 1.0 at a pixel.

Min-max normalization. A_{pm} and A_{appr} have quite different scales, e.g., $A_{pm} = \{0.0, \dots, 10^{-3}\}$ and $A_{appr} = \{0.4, \dots, 1.0\}$ in our experiments. To fuse two affinities, we apply Min-max normalization to them as follows:

$$\bar{A}^{(i,j)} = \frac{A^{(i,j)} - \min_{\substack{1\leq i\leq N\\1\leq j\leq M}} A^{(i,j)}}{\max_{\substack{1\leq i\leq N\\1\leq j\leq M}} A^{(i,j)} - \min_{\substack{1\leq i\leq N\\1\leq j\leq M}} A^{(i,j)}}. \tag{5}$$

Thus, we propose a simple affinity fusion model as follows:

$$Cost(\mathbf{x_{t|t-1}^i}, \mathbf{z_t^j}) = -\alpha \cdot \ln \bar{A}_{pm}^{(i,j)} \bar{A}_{appr}^{(i,j)}, \tag{6}$$

where α is a scale factor empirically set to 100. If one of affinities is close to zero value like 10^{-39}, the cost is set to 10000 to avoid that final cost becomes infinity value. Then, the final costs ranges 0 to 10000.

2.4. Mask merging

As shown in Mask merging module in Figure 2, we utilize segment mask based IoU (Mask IoU) measure which can calculate 2D pixel-wise overlapping ratio between two objects. Conventional 2D box based measure intersection-over-union (IoU) and Mask IoU are represented by:

$$IoU_{AB} = \frac{bbox(A) \cap bbox(B)}{bbox(A) \cup bbox(B)}, \tag{7}$$

$$Mask\ IoU_{AB} = \frac{mask(A) \cap mask(B)}{mask(A) \cup mask(B)}. \tag{8}$$

3. Experiments

GMPHD_SAF is evaluated on MOTSChallenge and KITTI-MOTS [18]. Inputs are image sequences and instance segmentation results created by Mask R-CNN X152 of Detectron2 [20]. We uniformly truncate detection results under threshold values that are 0.6 for cars and 0.7 for pedestrians. All experiments are conducted on Intel i7-7700K CPU @ 4.20GHz and DDR4 32.0GB RAM without GPU-acceleration. In Table 1 and 2, our MOTS trackers from $p2$ to $p5$ show incremental improvements compared to baseline method $p1$ whenever adding the key modules

Trackers	MOTSChallenge Test Sequences				KITTI-MOTS Test Sequences							
	Pedestrians				Cars				Pedestrians			
	sMOTSA↑	MOTSA↑	MOTSP↑	IDS↓	sMOTSA↑	MOTSA↑	MOTSP↑	IDS↓	sMOTSA↑	MOTSA↑	MOTSP↑	IDS↓
Track R-CNN [18]	40.6	55.2	76.1	576	67.0	79.6	85.1	692	47.3	66.1	74.6	481
MOTSFusion [11]	-	-	-	-	75.0	84.1	**89.3**	**201**	58.7	72.9	81.5	**279**
Ours ($p5$)	**61.8**	**73.4**	**84.8**	**524**	**75.4**	**86.7**	87.5	549	**62.8**	**78.2**	**81.6**	474
Ours ($p5^*$)	68.4	82.6	83.0	569	← Scene-specific truncation was used. (0.2 for MOTS20-07 and 0.6 for others)							

Table 3. **Evaluation results on MOTSChallenge and KITTI-MOTS test sets**.

"SAF in S2TA, Mask merging, and SAF in T2TA" one by one. In Table 3, our final model $p5$ not only achieves the best sMOTSA, MOSTA, and MOTSP scores against state-of-the-art methods [11, 18] but also runs at 11.4 FPS and 3.39 FPS speeds on KITTI-MOTS and MOTSChallenge.

4. Conclusion

In this paper, we propose a highly practical MOTS method named GMPHD_SAF which a feasible and easily reproducible combination of four key modules: GMPHD filter, Hierarchical data association, Mask merging, Simple affinity fusion. Those modules show incremental improvements in evaluation on training sets of KITTI-MOTS and MOTChallenge. Especially, ID-switch decreases by half compared to baseline method. In test sets of those two datasets, GMPHD_SAF achieves the best performance against the state-of-the-art MOTS methods.

Acknowledgements

This work was partly supported by Institute of Information Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2014-0-00077, Development of global multi-target tracking and event prediction techniques based on real-time large-scale video analysis).

References

[1] I. Reid S. Roth A. Milan, L. Leal-Taixé and K. Schindler. Motchallenge 2015.

[2] N. L. Baisa and A. Wallace. Development of a N-type GM-PHD filter for multiple target, multiple type visual tracking. *Journal of Visual Communication and Image Representation*, 59:257–271, 2019.

[3] Z. Fu, F. Angelini, J. Chambers, and S. M. Naqvi. Multi-level cooperative fusion of GM-PHD filters for online multiple human tracking. *IEEE Trans. Multimed.*, 21(9):2277–2291, 2019.

[4] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012.

[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *ICCV*, 2017.

[6] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(3):583–596, 2015.

[7] R. Jonker and A. Volgenant. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340, 1987.

[8] T. Kutschbach, E. Bochinski, V. Eiselein, and T. Sikora. Sequential sensor fusion combining probability hypothesis density and kernelized correlation filters for multi-object tracking in video data. In *AVSS*, 2017.

[9] I. Reid S. Roth L. Leal-Taixé, A. Milan and K. Schindler. MOT16: A benchmark for multi-object tracking. *arXiv Preprint arXiv:1603.00831*, 2016.

[10] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, 2019.

[11] J. Luiten, T. Fischer, and B. Leibe. Track to reconstruct and reconstruct to track. *arXiv Preprint arXiv:1910.00130*, 2019.

[12] J. Redemon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016.

[13] S. Ren, K. He, R. Girshick, , and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[14] S. Shi, X. Wang, and H. Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *CVPR*, 2019.

[15] Y. Song. Supplementary material for 'online multi-object tracking and segmentation with gmphd filter and simple affinity fusion", 2020. Supplied as additional material `bmtt2020_ymsong_suppl.pdf`, `https://github.com/SonginCV/GMPHD_SAF`.

[16] Y. Song, K. Yoon, Y. Yoon, K. C. Yow, and M. Jeon. Online multi-object tracking with GMPHD filter and occlusion group management. *IEEE Access*, 7:165103–165121, 2019.

[17] B.-N. Vo and W.-K. Ma. The Gaussian mixture probability hypothesis density filter. *IEEE Trans. Signal Process.*, 54(11):4091–4104, 2006.

[18] P. Voigtlaender, M. Krause, A. Ošep, J. Luiten, B. B. G. Sekar, A. Geiger, and B. Leibe. MOTS: Multi-object tracking and segmentation. In *CVPR*, 2019.

[19] X. Wang, M. Yang, S. Zhu, and Y. Lin. Regionlets for generic object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(10):2071–2084, Oct. 2015.

[20] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019.