Tracking Objects as Points

Xingyi Zhou, Vladlen Koltun, Philipp Krahenbuhl
UT Austin, Intel Labs

박태우

Introduction

- MOT(Multi Object Tracking)을 위한 tracker들이 Tracking-by-detection 구조를 이용했음.

- 최근 SOTA는 detection과 tracking의 기능을 분리시키지 않고 동시에 수행하도록 함.

- 본 논문은 CenterTrack이라는 detection과 tracking의 기능이 결합된 tracker를 제시.

- CenterTrack은 CenterNet이라는 detector의 input과 output에 추가적인 채널을 더해 tracking의 기능을 구현했기 때문에 point-based tracker임.

Introduction

CenterNet

Input : Image

Output : keypoint heatmap

- Bbox로 표현하던 detector와 달리 CenterNet은 object의 중심에 하나의 keypoint로 표현.

- Keypoint로부터 가로, 세로 길이를 통해 object의 크기 표현 가능.

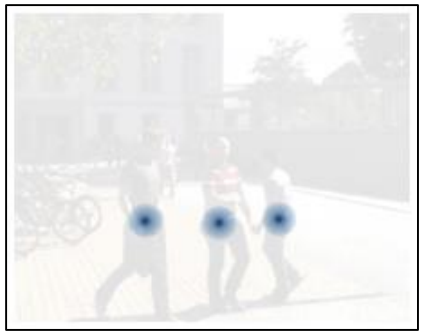
- Object의 위치를 keypoint로 표현하기 때문에 pose estimation에서 사용되는 기술과 유사함.

(CenterNet을 이용해 pose estimation으로 확장 가능)

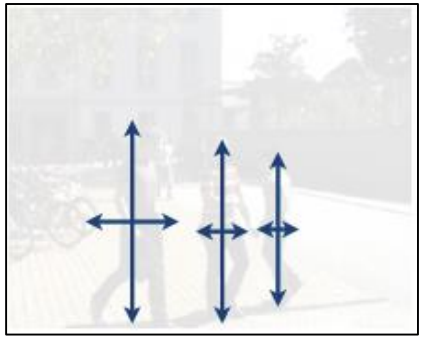
Architecture

Image $I^{(t)}$

CenterNet



Detection $\hat{Y}^{(t)}$

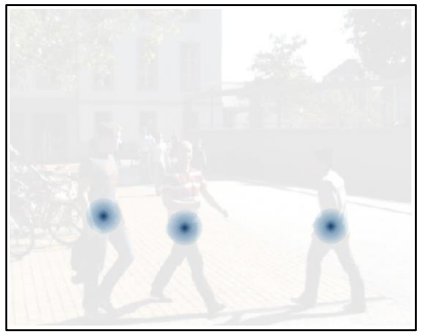


Size $\hat{S}^{(t)}$

Architecture

Image $I^{(t)}$

Image $I^{(t-1)}$



Tracklet $T^{(t-1)}$

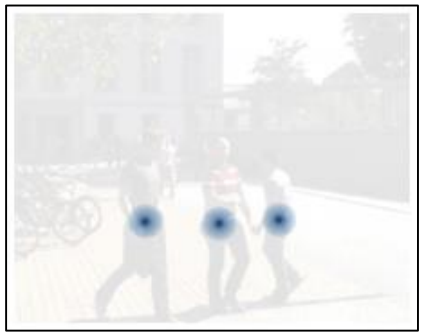
$\hat{Y}^{(t)}$: Heatmap

$\hat{S}^{(t)}$: Size map

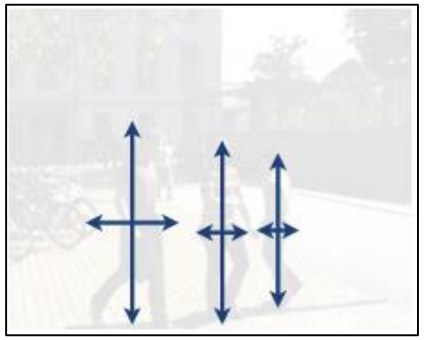
CenterTrack

$$T^{(t-1)} = \{b_0^{(t-1)}, b_1^{(t-1)}, \ldots\}_i$$

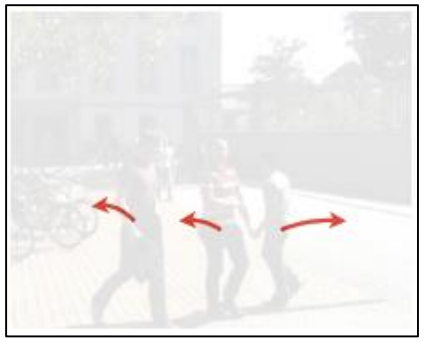
$$b = (\mathbf{p}, \mathbf{s}, w, id)$$



Detection $\hat{Y}^{(t)}$



Size $\hat{S}^{(t)}$



Offset $\hat{O}^{(t)}$

Heatmap, Size map, Offset map

Heatmap $\hat{Y}^{(t)}$, Size map $\hat{S}^{(t)}$, Offset map $\hat{S}^{(t)}$

네트워크의 output으로써 input 이미지 크기의 low-resolution 된 결과.

Heatmap의 경우 각 위치에서의 confidence scroe를 뜻하고 $\frac{W}{R}$ x $\frac{H}{R}$ x C 크기를 가짐.

Size map의 경우 해당 key point의 가로, 세로 길이를 뜻하고 $\frac{W}{R}$ x $\frac{H}{R}$ x 2 크기를 가짐.

Offset map의 경우 해당 key point의 t-1 frame과의 거리를 뜻하고 $\frac{W}{R}$ x $\frac{H}{R}$ x 2 크기를 가짐.

본 논문에서는 R = 4 로 지정.

Heatmap, Size map, Offset map

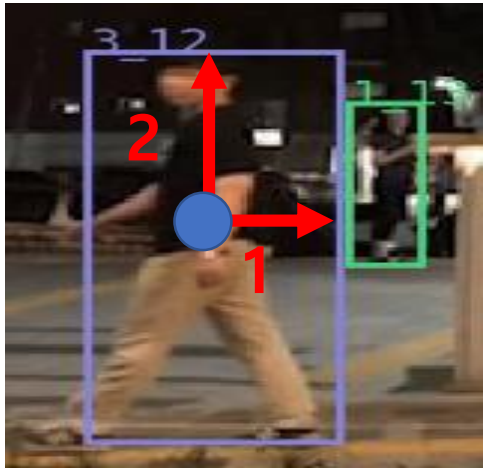


Image $I^{(t)}$

0	0	0	0	0
0	0	0	0	0
0	1	0	0	0
0	0	0	0	0
0	0	0	0	0

Heatmap
(C = person)

0	0	0	0	0
0	0	0	0	0
0	2	0	0	0
0	0	0	0	0
0	0	0	0	0

Size map 1

0	0	0	0	0
0	0	0	0	0
0	1	0	0	0
0	0	0	0	0
0	0	0	0	0

Size map 2

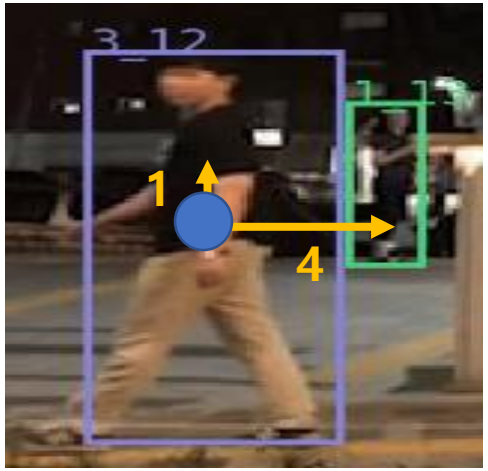


Image $I^{(t)}$

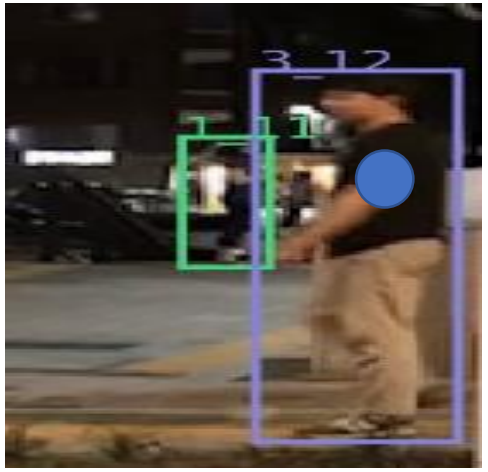


Image $I^{(t-1)}$

0	0	0	0	0
0	0	0	0	0
0	1	0	0	0
0	0	0	0	0
0	0	0	0	0

Offset map 1

0	0	0	0	0
0	0	0	0	0
0	4	0	0	0
0	0	0	0	0
0	0	0	0	0

Offset map 2

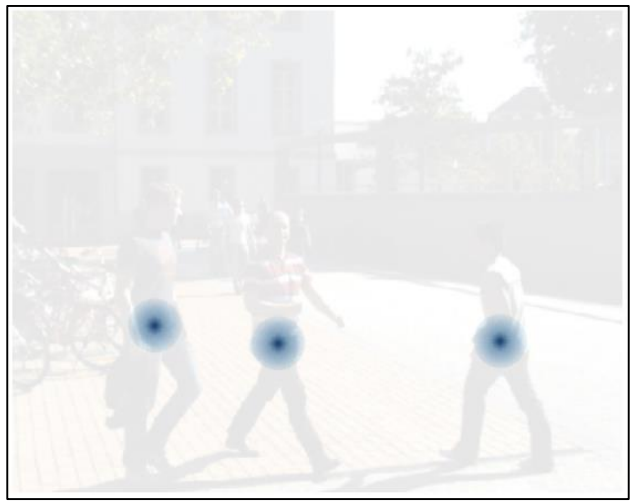
Tracking Algorithm

Tracking-conditioned detection

조금씩 가려져 있는 object의 경우에도 t-1 frame의 이미지를 input으로 넣기 때문에 정확도가 높아진다.

Single point로 표현되기 때문에 모든 detection 정보를 heatmap으로써 표현이 가능하다.

False positive를 줄이기 위해 confidence score가 τ 이상인 것들만 t-1 frame heatmap으로 만들어 input으로 사용.



Detection $\hat{Y}^{(t-1)}$

Tracking Algorithm

Association through offsets

- CenterTrack은 Offset map을 output으로 도출하는데 t에서 point와 t-1에서 point를 연결하기 위해 사용된다.

$$\hat{\mathbf{d}}^{(t)} = \hat{\mathbf{p}}^{(t)} - \hat{\mathbf{p}}^{(t-1)}$$

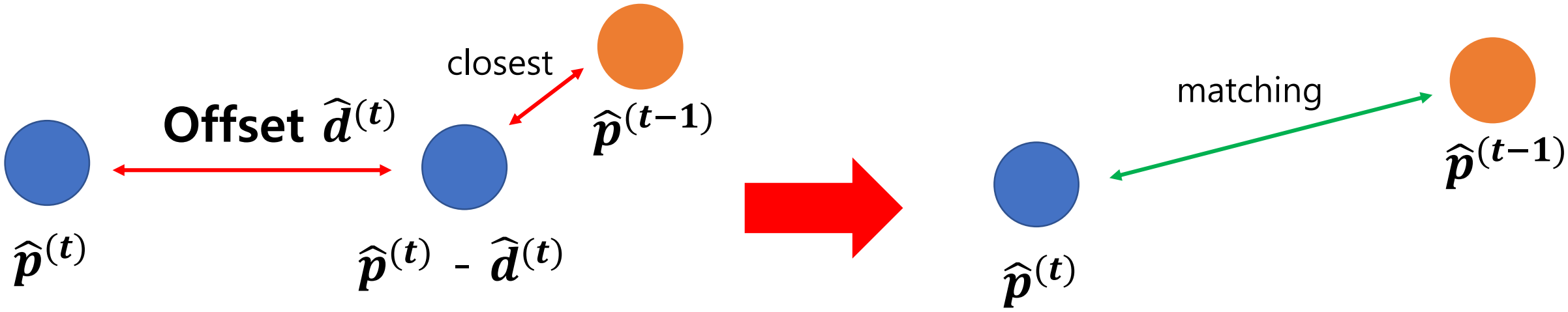
- 현재 frame t에서 예측한 $\hat{\mathbf{d}}^{(t)}$ 값이 특정 keypoint의 t-1 frame에서 위치와의 거리 차를 뜻한다.

- 즉, 현재 keypoint의 위치 $\hat{\mathbf{p}}^{(t)}$ 에서 $\hat{\mathbf{d}}^{(t)}$ 를 뺀 값을 t-1 frame에서 위치라고 가정하고 이 값과 가장 근접한 keypoint를 찾아 매칭시킨다.

- t에서 keypoint들과 t-1에서 keypoint의 모든 점을 비교하면 비교적 높은 정확도를 얻을 수 있지만, real time으로 진행되는 traker의 특성상 좋은 방법은 아니기 때문에 greedy matching algorithm을 사용한다.

Tracking Algorithm

Association through offsets



Result

	Time(ms)	MOTA ↑	IDF1 ↑	MT ↑	ML ↓	FP ↓	FN ↓	IDSW ↓
Tracktor17 [1]	666+D	53.5	52.3	19.5	36.6	12201	248047	2072
LSST17 [10]	666+D	54.7	**62.3**	20.4	40.1	26091	228434	**1243**
Tracktor v2 [1]	666+D	56.5	55.1	21.1	35.3	**8866**	235449	3763
GMOT	167+D	55.4	57.9	22.7	34.7	20608	229511	1403
Ours (Public)	**57+D**	**61.4**	53.3	**27.9**	**31.4**	15520	**196886**	5326
Ours (Private)	57	67.3	59.9	34.9	24.8	23031	158676	2898

Table 1: Evaluation on the MOT17 test sets (top: public detection; bottom: private detection). We compare to published entries on the leaderboard. The runtime is calculated from the HZ column on the leaderboard. +D means detection time, which is usually $>$ 100ms [31].

Result

	Time(ms)	MOTA ↑	MOTP ↑	MT ↑	ML ↓	IDSW ↓	FRAG ↓
AB3D [46]	4+D	83.84	85.24	66.92	11.38	**9**	**224**
BeyondPixel [35]	300+D	84.24	85.73	73.23	2.77	468	944
3DT [14]	30+D	84.52	85.64	73.38	2.77	377	847
mmMOT [54]	10+D	84.77	85.21	73.23	2.77	284	753
MOTSFusion [27]	440+D	84.83	85.21	3.08	2.77	275	759
MASS [18]	10+D	85.04	**85.53**	74.31	2.77	301	744
Ours	82	**89.44**	85.05	**82.31**	**2.31**	116	334

Table 2: Evaluation on the KITTI test set. We compare to all published entries on the leaderboard. Runtimes are from the leaderboard. +D means detection time.

	Time(ms)	AMOTA@0.2 ↑	AMOTA@1 ↑	AMOTP ↓
Mapillary [38]+AB3D [46]	-	6.9	1.8	1.8
Ours	45	**27.8**	**4.6**	**1.5**

Table 3: Evaluation on the nuScenes testing set. We compare to the official monocular 3D tracking baseline, which applies a state-of-the-art 3D tracker [46]. We list the average AMOTA@0.2, AMOTA@1, and AMOTP over all 7 categories.

Result

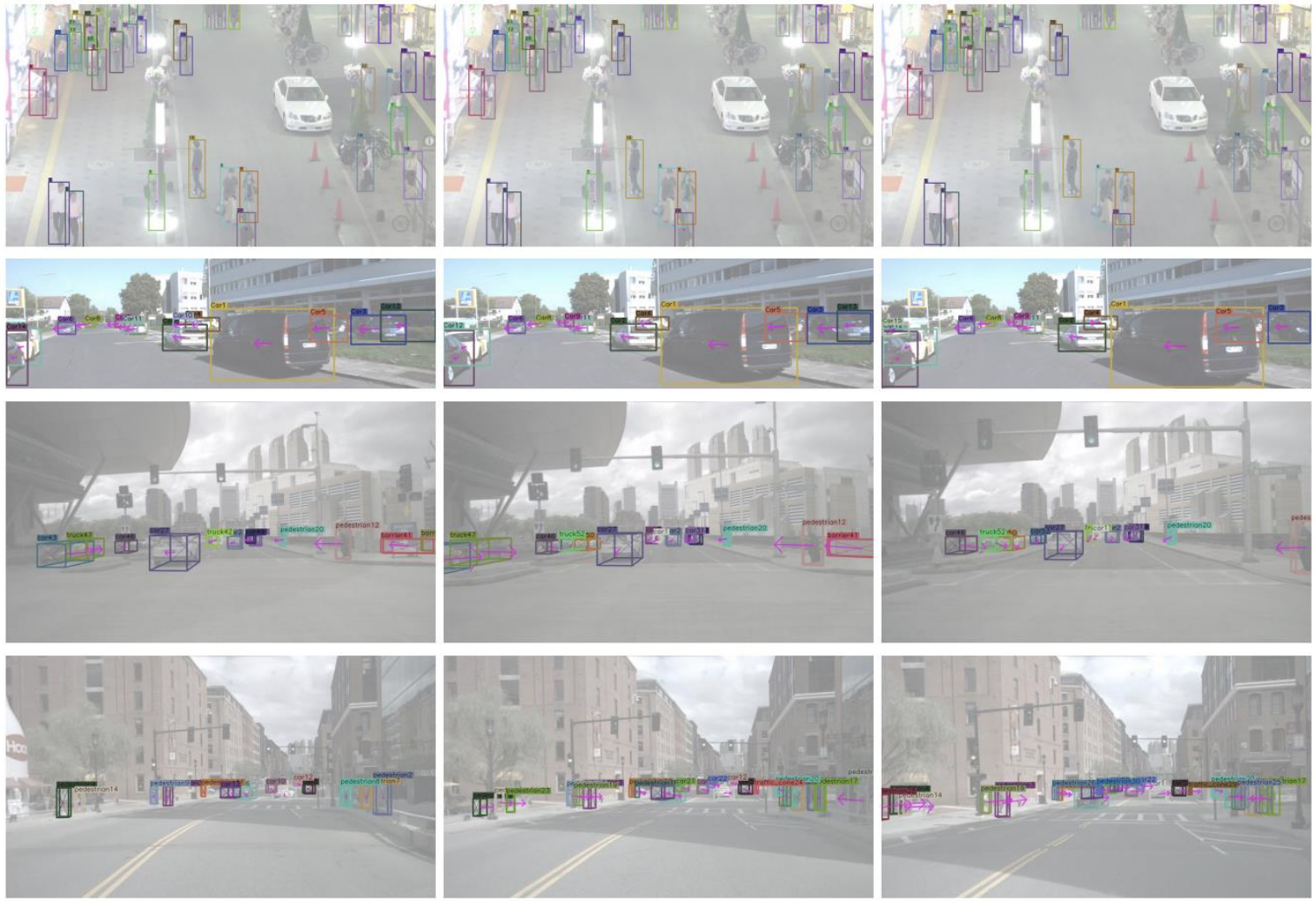


Fig. 3: Qualitative results on MOT (1st row), KITTI (2nd row), and nuScenes (3th and 4th rows). Each row shows three consecutive frames. We show the predicted tracking offset in arrow. Tracks are coded by color. Best viewed on the screen.

Reference

https://nuggy875.tistory.com/34