Supplementary Material on

Instance-Level Future Motion Estimation in a Single Image Based on Ordinal Regression

Kyung-Rae Kim[1] Whan Choi[1] Yeong Jun Koh[2] Seong-Gyun Jeong[3] Chang-Su Kim[1]
[1]Korea University [2]Chungnam National University [3]CODE42.ai

FM dataset: The table below lists the number of instances in each class in the FM dataset.

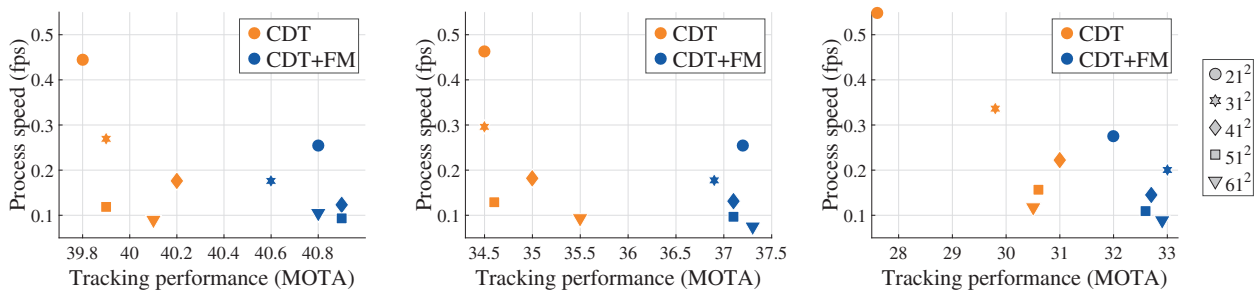
Type	Class	# Instances
Direction	N	1,012
	NE	1,062
	E	2,116
	SE	1,421
	S	914
	SW	1,125
	W	2,084
	NW	1,608
Velocity	Stop	1,886
	Slow	7,102
	Fast	2,354
Action	Sidewalk	6,219
	Crosswalk	2,108
	Jaywalk	3,015
Total		11,342

More FM estimation results: Figures S-3 and S-4 show examples of FM results on the FM dataset. Figures S-5 and S-6 provide examples of FM results of cars and animals, respectively.

Multi object tracking comparisons: Table S-1 compares CDT+FM with the baseline CDT. At the same number of search points, CDT+FM yields a slower processing speed, since it should perform the background motion compensation. However, by reducing the search region based on the FM, it provides more accurate tracking results, yielding a higher MOTA score. In Figure S-1, we plot MOTA scores versus processing speeds. At similar processing speeds, CDT+FM provides a significantly higher MOTA score than the baseline CDT.

Table S-1. Comparison of CDT+FM with the baseline CDT on the MOT17 dataset at low video frame rates. The reported fps scores are the processing speeds in frames per second.

Frame rate	# Samples	21^2		31^2		41^2		51^2		61^2	
	Method	MOTA	fps	MOTA	fps	MOTA	fps	MOTA	fps	MOTA	fps
5 fps	CDT	39.8	**0.445**	39.9	**0.270**	40.2	**0.176**	39.9	**0.118**	40.1	**0.090**
	CDT+FM	**40.8**	0.255	**40.6**	0.176	**40.9**	0.123	**40.9**	0.093	**40.8**	0.104
2 fps	CDT	34.5	**0.463**	34.5	**0.300**	35.0	**0.182**	34.6	**0.128**	35.5	**0.094**
	CDT+FM	**37.2**	0.255	**36.9**	0.178	**37.1**	0.131	**37.1**	0.097	**37.3**	0.075
1 fps	CDT	27.6	**0.548**	29.8	**0.337**	31.0	**0.222**	30.6	**0.157**	30.5	**0.117**
	CDT+FM	**32.0**	0.275	**33.0**	0.200	**32.7**	0.145	**32.6**	0.110	**32.9**	0.089



(a) (b) (c)

Figure S-1. MOTA scores versus tracking speeds on the MOT17 dataset at video frame rates of (a) 5 fps, (b) 2 fps, and (c) 1 fps.

Crowd Analysis:

We apply the proposed FM estimation algorithm to crowd analysis in a single image. By employing the estimated direction of each pedestrian in a crowd, we partition the crowd into several clusters and predict the group direction of each cluster. For the clustering, we use the simple k-means algorithm. To compute the distance between two instances, we use the weighted distance $D = D_{\mathrm{Euc}} + \lambda D_{\mathrm{FM}}$, where D_{Euc} is the Euclidean distance between the instances and D_{FM} is the cyclic difference of the directional indices. For example, D_{FM} between adjacent directions is 1, and the maximum D_{FM} is 4 for opposite directions. Also, $\lambda = 40$ is a weight parameter. After the clustering, we obtain the group direction of each cluster. To this end, for each direction, we compute the sum of the directional probabilities of instances within a cluster. Then, we select the direction, whose sum of the probabilities is maximal, as the group direction.

We capture various crowded scenes using a surveillance camera and detect pedestrians using the YOLOv3 detector [44]. Figure S-2 shows some of the crowd analysis results. In each column in Figure S-2, the top image shows predicted directions of pedestrians. Even in a single crowded scene, we see that the directions are predicted faithfully. By employing the directional information, the clustering is performed with the number of clusters $k = 5$ in the bottom image. Note that the bottom image is easier to understand than the top image, since it conveys information compactly through the data clustering. However, the grouping is not perfect. In Figure S-2(b), group 2 contains one pedestrian who is far away from the other three pedestrians in the same group. Also, one person who should be in group 2 is declared to belong to group 1. A more sophisticated algorithm than the k-means is required for this example. Figures S-7, S-8, and S-9 are more crowd analysis results.

(a) (b) (c)

Figure S-2. Crowd analysis based on the FM: In each column, the top image shows predicted directions of instances, and the bottom one is the corresponding clustering result with $k = 5$. If a cluster contains only one pedestrian, its label is not visualized.

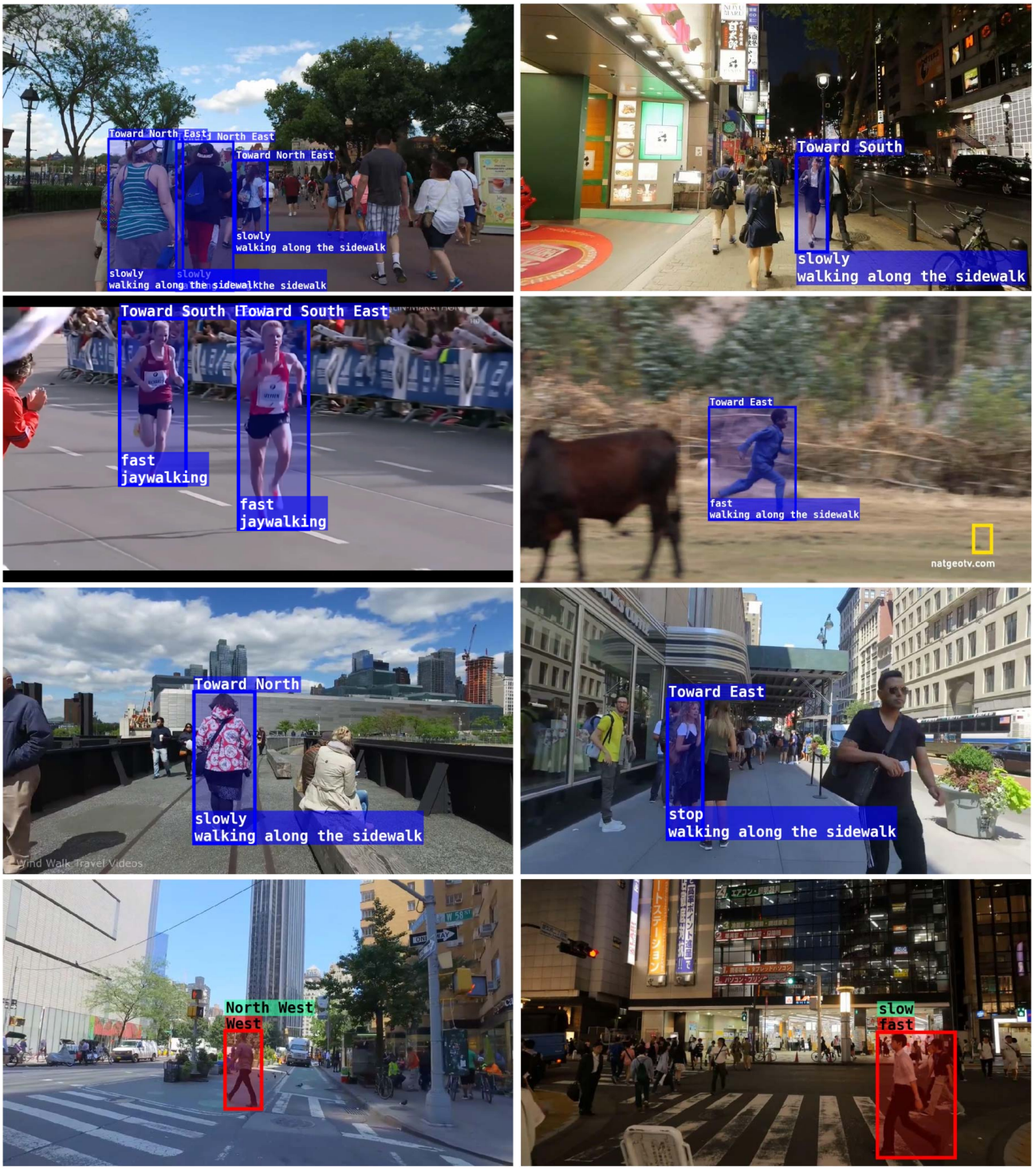


Figure S-3. FM estimation results on YouTube images in the FM dataset. The top three rows present correct results. On the other hand, the last row shows failure cases, where green labels are the ground-truth and red labels are predicted classes.

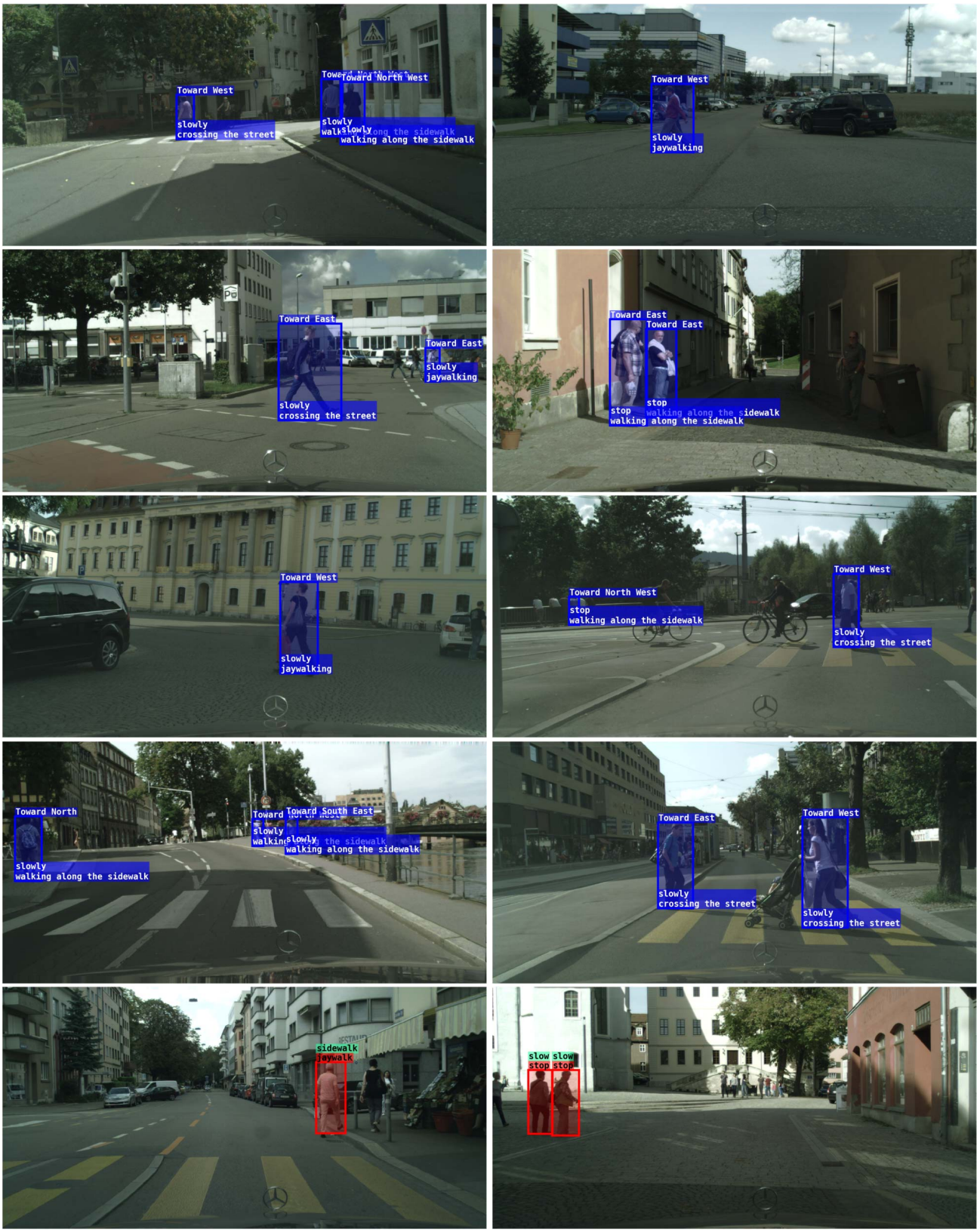


Figure S-4. FM estimation results on CityPerson images in the FM dataset. The top four rows present correct results. On the other hand, the last row shows failure cases, where green labels are the ground-truth and red labels are predicted classes.

Figure S-5. FM estimation results of cars. The top five rows present correct results. On the other hand, the last two rows show failure cases, where green labels are the ground-truth and red labels are predicted classes.

Figure S-6. FM estimation results of animals. The top four rows present correct results. On the other hand, the last two rows show failure cases, where green labels are the ground-truth and red labels are predicted classes.

Figure S-7. Crowd analysis based on the FM etimation.

Figure S-8. Crowd analysis based on the FM estimation.

Figure S-9. Crowd analysis based on the FM estimation.