D^3: Duplicate Detection Decontaminator for Multi-Athlete Tracking in Sports Videos

Rui He, Zehua Fu, Qingjie Liu, Yunhong Wang, Xunxun Chen

Supplementary Materials

1 Explaination of setting limitation K in RH

Here we utilize functional analysis to explain why the limitation K in RH can be set.

Firstly, we try to construct a metric space. A metric space is a pair (X, d) which consists of a nonempty set X together with a metric or distance function d. Suppose that X is the set of bounding boxes of all objects in a video. For $x, y \in X$, we define a metric $d : X \times X \to \mathbb{R}$ by writing

$$d(x,y) = \begin{cases} 0, & x = y, \text{ or } IoU(x,y) = 1; \\ o(\alpha), & 0 < IoU(x,y) < 1, \text{ for each } \epsilon > 0, 0 < o(\alpha) < \epsilon, \\ & \text{or } d(x,z) = o(\alpha) \text{ and } d(y,z) = o(\alpha) \\ |x-y|, & IoU(x,y) = 0. \end{cases}$$

where $o(\alpha)$ is an infinitesimal, and $IoU(x, y) > 0$ means x intersects y and $IoU(x, y) = 0$ means not. If X is a metric space, d should satisfy the three following conditions:

$(M1)$(Nonnegativity)$d(x, y) \geq 0$ for all $x, y \in X$, and $d(x, y) = 0$ if and only if $x = y$;

$(M2)$(Symmetry)$d(x, y) = d(y, x)$, for every $x, y \in X$;

$(M3)$(Triangle inequality)for every $x, y, z \in X$, we have $d(x, z) \leq d(x, y) + d(y, z)$.

Proof. It is easy to see that d satisfies $(M1)$ and $(M2)$. To check $(M3)$, when $0 < IoU(x, z) < 1$, for every $y \in X$ and $y \neq x, z$, it is easy to check $d(x, y) + d(y, z) \geq 2o(\alpha) > o(\alpha) = d(x, z)$. When $IoU(x, z) = 0$, if $0 < IoU(x, y) < 1$ and $0 < IoU(y, z) < 1$, according to the defination of $d(x, y)$, even $IoU(x, z) = 0$, $d(x, z) = o(\alpha) \leq d(x, y) + d(y, z)$; If $0 < IoU(x, y) < 1$ and $IoU(y, z) = 0$, $d(x, y) = o(\alpha)$, $d(y, z) = |y - z|$, thus

$$d(x,y) + d(y,z) = o(\alpha) + d(y,z) > d(x,x) + d(x,z) = d(x,z).$$

if $IoU(x, y) = 0$ and $0 < IoU(y, z) < 1$, it is the same as above formular; If $IoU(x, y) = 0$ and $IoU(y, z) = 0$, it means that x,y,z have no intersection with each other, according to a character of absolute value inequality that

$$d(x,y) + d(y,z) = |x-y| + |y-z| \geq |x-y+y-z| = |x-z| = d(x,z).$$

Thus $(M3)$ is also satisfied that (X, d) is a metric space.

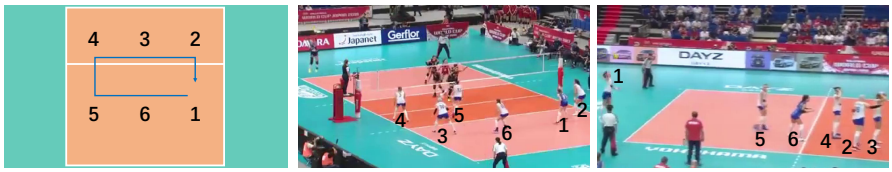
Secondly, we construct Cauchy sequences. A sequence in (X, d) is said to be a Cauchy sequence, if $d(x_n, x_m) \to 0$ as $n, m \to \infty$, i.e., for every $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that $d(x_n, x_m) < \epsilon$ for all $n, m > N$.

During a rally of a volleyball match, we regard the trajactory of an individual as a sequence $\{x_n\}$ in (X, d). A rally will not stop making that $n, m \to \infty$, and the individual will move slower and slower leading to $d(x_n, x_m) \to 0$. Thus the trajactory of an individual $\{x_n\}$ is a Cauchy sequence.

Finally, we proof the Cauchy sequences has a limit. If we regard the volleyball court as x, according to the conditions above, we will get $d(x_n, x) \to 0$ as $n \to \infty$, which means that the limit of $\{x_n\}$ could be x. But $\{x_n\}$ is not convergent for the possible limit $x \notin X$ so that the metric space (X, d) is not complete.

However, that the limit exists is sufficient to explain why we can set a limitation K in RH. As we said before, the trajactory of an individual $\{x_n\}$ has a limit means $\{x_n\}$ is 'converging' to the court. That is to say, in a volleyball match all the trajactories of individuals are 'converging' to the court. If there are no additional trajactories or the original ones does not reduce, the number of the trajectories will be fixed during tracking. It is very unlike pedestrian video in which the the trajactories of pedestrians are not 'converging'. Thus in each frame we could set a limitation K as the total number of owners of the trajectories, such as $K = 12$ in volleyball, $K = 10$ in basketball, $K = 22$ in soccer, according to the rules of different team sports.

2 Method of labeling according to volleyball rules



(a) (b) (c)

Fig. 1. Illustration of position and rotate: (a) position and rotate; (b) frame 503 in test_0120001; (c) frame 822 in test_0120001.

In our dataset, videos of broadcast and fixed view are all collected. The fixed ones are easy to label, but the others are difficult because of shot changing. Thus we labeled the annotation according to the rules of volleyball game in [1]. We mainly briefly introduce two rules, positions and rotation, which we use to label the broadcast view videos.

As shown in Figure 1(a), the player position numbers are as follows: three players along the net are front-row players and occupy positions 4 (front-left), 3

(front-centre) and 2 (front-right); the other three are back-row players occupying positions 5 (back-left), 6 (back-centre) and 1 (back-right). Relative positions between players: each front-row player must have at least a part of his/her foot closer to the centre line than the feet of the corresponding back-row player; each right (left) side player must have at least a part of his/her foot closer to the right (left) sideline than the feet of the centre player in that row. Figure 1(b) displays an example of correct positions. Position 5 is only corresponding to Position 4 and Position 6 and not closer to the centre line than Position 4 or the right sideline than Position 6.

In Figure 1(b), the black team is serving. If the white team scores, players in white gained the right to serve and should rotate one position clock-wise : the player in position 2 rotates to position 1 to serve, the player in position 1 rotates to position 6, etc, as the blue arc arrow shown in Figure 1(a) and an instance in Figure 1(c). As a result, we had to label the tracking annotations according to volleyball game rules, rather than adjust the boxes in each frame in the usual way. In real broadcast view, the same player's jersey number, often invisible, and low-resolution video make MOT in sports video becoming a challenging task.

3 More methods compared on RallyTrack

We provide another comparison to FairMOT [3] as shown in Table 1. Compared to TransTrack, FairMOT shows a different performance on RallyTrack. The reason may be that FairMOT has a very different tracking mechanism which is based on the ReID theory. So FairMOT is good at data association with IDF1 42.0 and AssA 26.0 rather than the detection which has FN 26002. However, D^3 is a transformer-based method that mainly solves detection problems. This phenomenon will encourage us to explore the probability of extending D^3 to more methods.

Table 1. Experiments on RallyTrack. Our method makes an amazing 9.2 promotion on MOTA, 7.0 on IDF1 and 4.5 on HOTA to baseline TransTrack (TT). Best in bold.

Model	MOTA↑	IDF1↑	MOTP↑	MT↑	FP↓	FN↓	IDS↓	HOTA↑	DetA↑	AssA↑
TT [2]	59.5	28.8	77.8	70.6	15370	19489	2049	27.9	51.9	15.2
FairMOT [3]	62.4	42.0	73.6	45.2	6805	26002	1464	35.4	49.1	26.0
TT+RH	62.0	33.3	77.8	66.7	12557	20310	**1788**	30.2	52.3	17.7
TT+D^3	66.4	29.7	78.1	**78.6**	13676	**14848**	2107	29.2	55.8	15.5
TT+D^3+RH	**68.7**	**35.8**	**78.1**	77.0	**11359**	15350	1847	**32.4**	**56.3**	**18.9**

4 Different Lower Bounds on MOT17, MOT20

We provide detailed experiments on MOT17, MOT20 to show the different Lower Bound settings in Table 2. The hyperparameters in the first column mean Lower Bound (LB) in D^3. LB is chosen according to different self-GIoU losses from different datasets. Different self-GIoU losses are caused by different resolutions of videos. From our observation, the resolutions of videos in RallyTrack, MOT17, MOT16, DanceTrack are of slight difference, so the LBs are closer to each other. Meanwhile, the resolution of videos in MOT20 is much larger that the others, so the LB is as well higher than the others.

Table 2. Different Lower Bounds on MOT17, MOT20.

D^3	Epoch	MOTA↑	IDF1↑	MOTP↑	MT↑	FP↓	FN↓	IDS↓	HOTA↑	DetA↑	AssA↑
w/o	150	65.1	63.6	81.9	36.8	1918	16440	**438**	52.6	54.0	51.7
0.009	150	64.5	62.4	82.1	36.0	2069	16457	581	51.8	54.0	50.2
0.010	150	65.3	62.9	82.2	38.3	1849	**16358**	457	53.0	54.4	52.1
0.011	150	64.2	63.9	82.0	37.2	2091	16700	520	53.1	53.7	52.9
w/o	30	64.9	62.6	82.0	36.3	1862	16537	477	52.1	53.9	50.7
0.010	**30**	**65.3**	**63.6**	**82.2**	**38.6**	**1833**	16398	480	**53.4**	**54.5**	**52.8**
mot20											
w/o	30	72.5	63.2	82.9	51.6	12882	153K	2978	52.4	59.4	46.3
0.015	30	72.8	63.3	82.9	52.9	13935	151K	2873	52.5	59.8	46.2
0.017	**30**	**73.2**	**64.6**	82.9	**53.4**	**12831**	**149K**	**2808**	**53.6**	**60.1**	**47.9**
0.019	30	72.7	63.7	82.9	52.3	13523	151K	3013	52.7	59.8	46.6

References

1. FIVB: Official volleyball rules 2017-2020. `http://www.fivb.org/EN/Refereeing-Rules/RulesOfTheGame_VB.asp` (2016)
2. Sun, P., Jiang, Y., Zhang, R., Xie, E., Cao, J., Hu, X., Kong, T., Yuan, Z., Wang, C., Luo, P.: Transtrack: Multiple-object tracking with transformer. arxiv **abs/2012.15460** (2020)
3. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: On the fairness of detection and re-identification in multiple object tracking. Int. J. Comput. Vis. **129**(11), 3069–3087 (2021)