# [2017] Simple Online and Realtime Tracking with a Deep Association Metric

2021年5月25日 16:33

Tags：多目标跟踪；实时；单摄像头；DeepSORT

## 基本信息

```
@inproceedings{wojke2017simple,
  title={Simple online and realtime tracking with a deep association metric},
  author={Wojke, Nicolai and Bewley, Alex and Paulus, Dietrich},
  affiliation={University of Koblenz-Landau, Queensland University of Technology},
  booktitle={2017 IEEE international conference on image processing (ICIP)},
  pages={3645--3649},
  year={2017},
  organization={IEEE}
}
```

## 主要贡献

1. 针对 Sort 算法容易出现跟踪失败，频繁切换标签的问题，引入了 外观特征，使得在较长时间的生命周期中，当目标对象受到遮挡，也能有较好的跟踪性能，有效缩减标签切换频次。

## 主要方法

### Track Handling and State Estimation

假设：摄像头是校准的，不会自移动

#### 状态估计

1. 目的：预测对象在下一帧的位置信息。
2. 主要方法：Kalman filtering 卡尔曼滤波
3. 跟踪场景的定义：8 维的状态空间 $(\mathrm{u}, \mathrm{v}, \gamma, h, \dot{x}, \dot{y}, \dot{\gamma}, \dot{h})$
   边界框的中心位置 $(\mathrm{u}, \mathrm{v})$， 长宽比 $\gamma$，高度 $h$，以及它们各自的速度
4. 模型：标准的 kalman filter（匀速运动模型）+ 线性观测模型
5. 说明：当检测与轨迹关联时，使用预测值与观测值进行最优估计，更新目标状态；否则，使用线性匀速模型更新状态

#### 轨迹处理

1. 目的：确定什么时候产生新的轨迹，什么时候轨迹终止
2. 方法：
   (1) 轨迹消亡：设定最大生命周期阈值：$\mathrm{A}_{max}$。对于任一 track $\mathrm{k}$，从最后一次被成功关联帧开始计数，记为 $\mathrm{a}_\mathrm{k}$。$\mathrm{a}_\mathrm{k}$ 在 Kalman filter预测时增加，当该 track 被成功关联时 置 $\mathrm{a}_\mathrm{k} = 0$。当 $\mathrm{a}_\mathrm{k}$ 大于阈值 $\mathrm{A}_{max}$，则认为轨迹离开消失，从 track set 中删除该轨迹信息
   (2) 轨迹产生：出现无法与已有track 匹配的新 track 时，这些无法关联的 track 将在它们存在的前三帧中被认为是 暂时的。在此期间，若能成功关联，则认为是已有 track 的更新，否则，认为产生了一个新的 track。

## Assignment Problem

1. 目的：将预测的目标状态 与 已有的目标 进行关联操作，融合 动作信息 和 外观信息
2. 方法：匈牙利算法（Hungarian algorithm），分配问题
3. 代价矩阵：
   (1) 位置关联度：基于Kalman预测位置与检测位置，马氏距离
   $\mathbf{d}^{(1)}(\mathbf{i},\mathbf{j}) = \left(\mathbf{d_j} - \mathbf{y_i}\right)^{\mathbf{T}} \mathbf{S_i^{-1}} \left(\mathbf{d_j} - \mathbf{y_i}\right)$
   $\mathbf{b}_{\mathbf{i,j}}^{(1)} = \mathbf{1}\left[\mathbf{d}^{(1)}(\mathbf{i},\mathbf{j}) \le \mathbf{t}^{(1)}\right]$
   设定阈值 $\mathrm{t}^{(1)}$，基于卡方分布 0.95 分位点，对于 4 个自由度，$\mathrm{t}^{(1)} = 9.4877$
   (2) 外观关联度：深度神经网络抽取，余弦距离
   与最后关联的 $\mathrm{L_k} = 100$ 个track的外观进行匹配
   $\mathbf{d}^{(2)}(\mathbf{i},\mathbf{j}) = \mathbf{min}\left\{\mathbf{1} - \mathbf{r_j^T r_k^{(i)}} \middle| \mathbf{r_k^{(i)}} \in \mathbf{R_i}\right\}$，$\|\mathrm{r_j}\| = 1$
   $\mathbf{b}_{\mathbf{i,j}}^{(2)} = \mathbf{1}\left[\mathbf{d}^{(2)}(\mathbf{i},\mathbf{j}) \le \mathbf{t}^{(2)}\right]$
   (3) 信息融合：马氏距离提供短时的位置信息，余弦距离提供长时的外观信息
   $\mathrm{c_{i,j}} = \lambda \mathrm{d}^{(1)}(i,j) + (1-\lambda) d^{(2)}(i,j)$
   $\mathrm{b}_{i,j} = \prod_{m=1}^{2} b_{i,j}^{m}$
   (4) 当镜头存在大量运动时，设置 λ = 0 是合理的

## Matching Cascade

1. 全局最优匹配存在的问题：当一个 track 被长时间遮挡时，基于 Kalman Filter 的位置预测的不确定性将会增加。但当两个 track 匹配同一个 detection 时，马氏距离会更倾向于具有较大不确定性的track。这将导致跟踪更容易被中断以及不稳定的跟踪。
2. 级联匹配 Matching Cascade

**Listing 1** Matching Cascade

**Input:** Track indices $\mathcal{T} = \{1,\dots,N\}$, Detection indices $\mathcal{D} = \{1,\dots,M\}$, Maximum age $A_{\max}$

Compute cost matrix $\boldsymbol{C} = [c_{i,j}]$ using Eq. 5
Compute gate matrix $\boldsymbol{B} = [b_{i,j}]$ using Eq. 6
Initialize set of matches $\mathcal{M} \leftarrow \emptyset$
Initialize set of unmatched detections $\mathcal{U} \leftarrow \mathcal{D}$
**for** $n \in \{1,\dots,A_{\max}\}$ **do**
  Select tracks by age $\mathcal{T}_n \leftarrow \{i \in \mathcal{T} \mid a_i = n\}$
  $[x_{i,j}] \leftarrow \text{min\_cost\_matching}(\boldsymbol{C}, \mathcal{T}_n, \mathcal{U})$
  $\mathcal{M} \leftarrow \mathcal{M} \cup \{(i,j) \mid b_{i,j} \cdot x_{i,j} > 0\}$
  $\mathcal{U} \leftarrow \mathcal{U} \setminus \{j \mid \sum_i b_{i,j} \cdot x_{i,j} > 0\}$
**end for**
**return** $\mathcal{M}, \mathcal{U}$

3. 匹配算法描述：
   (1) 对于 confirmed track，使用 级联匹配
   (2) 对于 unconfirmed track & unmatched track，使用 IOU 匹配，解决由于遮挡引起的外观急剧变化 造成的影响

## Deep Appearance Descriptor

1. 网络结构：

| Name | Patch Size/Stride | Output Size |
|---|---|---|
| Conv 1 | $3 \times 3/1$ | $32 \times 128 \times 64$ |
| Conv 2 | $3 \times 3/1$ | $32 \times 128 \times 64$ |
| Max Pool 3 | $3 \times 3/2$ | $32 \times 64 \times 32$ |
| Residual 4 | $3 \times 3/1$ | $32 \times 64 \times 32$ |
| Residual 5 | $3 \times 3/1$ | $32 \times 64 \times 32$ |
| Residual 6 | $3 \times 3/2$ | $64 \times 32 \times 16$ |
| Residual 7 | $3 \times 3/1$ | $64 \times 32 \times 16$ |
| Residual 8 | $3 \times 3/2$ | $128 \times 16 \times 8$ |
| Residual 9 | $3 \times 3/1$ | $128 \times 16 \times 8$ |
| Dense 10 | | 128 |
| Batch and $\ell_2$ normalization | | 128 |

2. 训练数据集：基于 大规模 person re-identification dataset，超过 1,100,000张图片，以及1,261个行人

## 实验结果

| | | MOTA ↑ | MOTP ↑ | MT ↑ | ML ↓ | ID ↓ | FM ↓ | FP ↓ | FN ↓ | Runtime ↑ |
|---|---|---|---|---|---|---|---|---|---|---|
| KDNT [16]* | BATCH | 68.2 | 79.4 | 41.0% | 19.0% | 933 | 1093 | 11479 | 45605 | 0.7 Hz |
| LMP_p [17]* | BATCH | **71.0** | **80.2** | **46.9%** | 21.9% | 434 | **587** | 7880 | **44564** | 0.5 Hz |
| MCMOT_HDM [18] | BATCH | 62.4 | 78.3 | 31.5% | 24.2% | 1394 | 1318 | 9855 | 57257 | 35 Hz |
| NOMTwSDP16 [19] | BATCH | 62.2 | 79.6 | 32.5% | 31.1% | **406** | 642 | **5119** | 63352 | 3 Hz |
| EAMTT [20] | **ONLINE** | 52.5 | 78.8 | 19.0% | 34.9% | 910 | **1321** | **4407** | 81223 | 12 Hz |
| POI [16]* | **ONLINE** | **66.1** | 79.5 | **34.0%** | 20.8% | 805 | 3093 | 5061 | **55914** | 10 Hz |
| SORT [12]* | **ONLINE** | 59.8 | **79.6** | 25.4% | 22.7% | 1423 | 1835 | 8698 | 63245 | **60 Hz** |
| Deep SORT (Ours)* | **ONLINE** | 61.4 | 79.1 | 32.8% | **18.2%** | **781** | 2008 | 12852 | 56668 | 40 Hz |

**Table 2**: Tracking results on the MOT16 [15] challenge. We compare to other published methods with non-standard detections. The full table of results can be found on the challenge website. Methods marked with * use detections provided by [16].