简介

作者提出了一种top-down多人姿态跟踪的方法。结合Single-person Pose Tracking(SPT)和(Visual Object Tracking)VOT两个module。然后利用一个SGCN(Siamese Graph Convolution Network)实现Re-ID，给每个entity分配一个ID实现多人追踪。
作者说这是第一篇实现在线top-down姿态追踪的工作。(hehe)
和之前的VOT工作不同，作者追踪的主要对象并不是bbox，而是pose，通过扩大当前帧pose roi区域，然后估计下一帧pose，而不是先追踪bbox再估计pose。所以其实结合了VOT和SPT。
主要贡献有三：

1. 提出了一种新的姿态追踪框架：VOT + SPT + ReID
2. 提出了用SGCN做ReID
3. 实验证明SOTA

方法

1. 粗略的位置估计可以通过SPE精炼成骨架姿态
2. 骨架位置可以用来指导得出粗略的人的位置
3. SPT的一种较好的策略就是着两者递归估计。

但是如果只是把MPT当作多个SPT，出现的问题是当SPT目标丢失之后，或者两个SPT相隔很近的时候，该怎么区分这两个SPT。所以作者提出用SPT + Pose match的方法。
首先下一帧人物的bounding box是通过当前预测的Pose得到的。找到人物的最小和最大坐标并且向外扩展20%以得到ROI区域。这个扩大的ROI区域就是下一帧姿态估计的预测区域。如果所估计姿态的平均分数小于阈值，那么就认为当前bounding box当中目标丢失。

$$\text{state} = \begin{cases} \text{tracked}, & \text{if } \bar{s} > \tau_s, \\ \text{lost}, & \text{otherwise.} \end{cases} \qquad (1)$$

目标丢失后有两种模式：

1. Fixed Keyframe Interval（FKI）等待下一个关键帧(key-frame)的出现然后再次进行detection并且根据tracking history分配ID或者重新产生ID。
2. Adaptive Keyframe Interval(AKI) 一旦有目标消失就马上进行detection和identity association。

FKI能保证帧率stable，而AKI的优点是对于non-complex的视频平均帧率更高。
作者用的策略是两者结合，准确率更高。

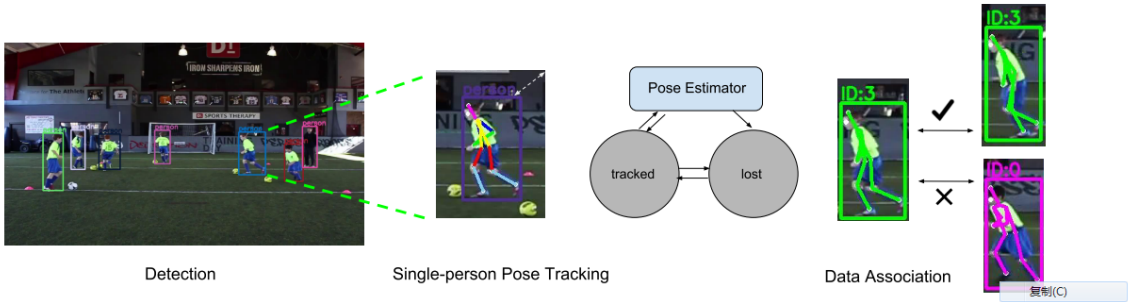

Figure 1. Overview of the proposed online pose tracking framework. We detect human candidates in the first frame, then track each candidate's position and pose by a single-person pose estimator. When a target is lost, we perform detection for this frame and data association with a graph convolution network for skeleton-based pose matching. We use skeleton-based pose matching because visually similar candidates with different identities may confuse visual classifiers. Extracting visual features can also be computationally expensive in an online tracking system. Pose matching is considered because we observe that in two adjacent frames, the location of a person may drift away due to sudden camera shift, but the human pose will stay almost the same as people usually cannot act that fast.

identity association

用spatial consistency和pose consistency来共同确定。

- spatial consistency: 如果相邻帧的bbox的IOU够高，那么就认为他们是同属于一个目标。

$$m(t_k, d_k) = \begin{cases} 1, & \text{if } o(t_k, \mathcal{D}_{i,k}) > \tau_o, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

  其中t是target，d是detection的结果，k是的k个key-frame。m代表matching flag。o代表max overlap ratio。
- pose consistency: 当出现目标比较大的漂移的时候，空间一致性假设就不成立，所以这里提出了一个Re-ID模块，但是作者觉得基于appearance的Re-ID太耗费时间，所以提出了基于graph的Re-ID，因为作者认为相邻帧就算人物位置发生很大的变化但是动作不可能变得太快。

Siamese Graph Convlutional Networks

作者首先将pose的2D坐标作为输入，构造了一个图模型，每个joints都是一个graph node，然后人体关节之间的连接作为graph edge。这个模型当做一个孪生网络，encode a pair of input keypoints，然后计算他们之间的距离。这个模型的训练loss为contrastive loss:

$$\mathcal{L}(p_j, p_k, y_{jk}) = \frac{1}{2} y_{jk} D^2 + \frac{1}{2}(1 - y_{jk}) \max(0, \epsilon - D^2), \tag{3}$$

其中 $D = ||f(p_j) - f(p_k)||_2$ 代表l2-norm normalized过的两个在隐空间的embedding距离。 $y_{jk} \in \{0, 1\}$ 代表pose j和pose k时候是同一个pose or not。而 ϵ 表示minimum distance margin

that pairs depicting different poses should satisfy。

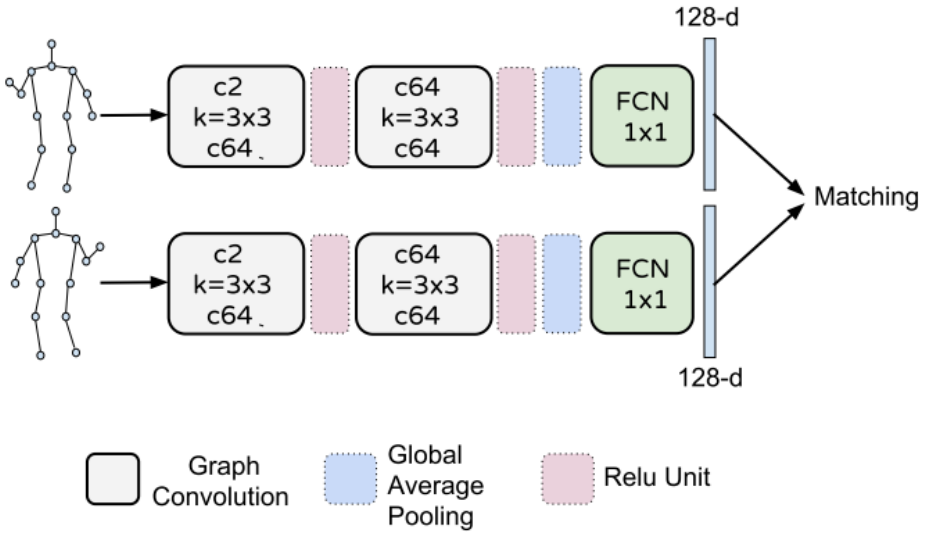

Figure 3. The siamese graph convolution network for pose matching. We extract two feature vectors from the input graph pair with shared network weight. The feature vectors inherently encode the spatial relationship among the human joints.

基于skeleton的图卷积

因为要保证卷积的输入数目保持固定所以作者用了一些tricks。

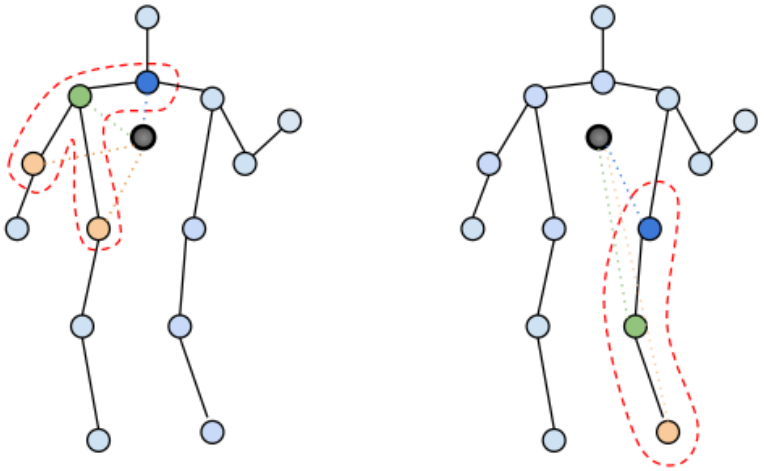

Figure 4. The spatial configuration partitioning strategy proposed in [38] for graph sampling and weighting to construct graph convolution operations. The nodes are labeled according to their distances to the skeleton gravity center (black circle) compared with that of the root node (green). Centripetal nodes have shorter distances (blue), while centrifugal nodes have longer distances (yellow) than the root node.

实验

数据集

实验选择的benchmark是PoseTrack数据集。目前包含593个训练视频，74个验证集视频，375个测试集视频。由于对于测试集有提交次数限制（每个任务4次），验证集没有限制，因此作者做ablation study。posetrack'18 测试集当时还没用公开，测试部分用的是posetrack'17测试集

评价标准

对于pose而言用的评价标准是mAP，对于追踪的评价标准是MOT。

实现细节

作者用的object detector是DCN pretrained版本。

- DCN + R-FCN + Resnet101
- DCN + FPN + Resnet101 （Fast R-CNN HEAD）

-	Method / Thresh	0.1	0.2	0.3	**0.4**	0.5
Prec	Deformable FPN	17.9	27.5	32.2	**34.2**	35.7
	Deformable R-FCN	15.4	21.1	25.9	30.3	34.5
Recall	Deform FPN	87.7	86.0	84.5	**83.0**	80.8
	Deform R-FCN	87.7	86.5	85.0	82.6	80.1

Table 2. Comparison of detectors: Precision-Recall on PoseTrack 2017 validation set. A bounding box is correct if its IoU with GT is above certain threshold, which is set to 0.4 for all experiments.

-	**Estimation (mAP)**			**Tracking (MOTA)**		
Method	Wri	Ankl	Total	Wri	Ankl	Total
GT Detections	74.7	75.4	**81.7**	56.3	56.2	**67.0**
Deform FPN-101	70.2	64.7	**74.6**	54.6	48.7	**61.3**
Deform RFCN-101	69.0	64.3	**73.7**	52.2	47.4	**59.0**

Table 3. Comparison of offline pose tracking results using various detectors on PoseTrack'17 validation set.

PoseTrack的数据集其实是没有bounding box的，作者采取的方式是求出min max坐标的bbox然后向外扩展20%。

SPE是从CPN101和MSRA152中选取。先在PoseTrack'17和COCO kp的混合数据集上训练了260个epoch然后再PoseTrack'17上finetune了40 epochs，主要是为了mitigate head和neck的回归准确。因为在COCO数据集当中。bottom-head和top-head位置没有给定，所以训练的时候是插值得到的。finetune的时候采用了online hard keypoint mining的策略，只关注15个关键点中7个hardest keypoints。

对于SGCN模块，2个GCN层和1个卷积层，contrastive loss，最后输出维度为128的vector。
作者自己做了一个SGCN的数据集，其中positive pair的选取是同一段视频同一个人在临近帧的pose。negative pair是同一段视频，不同人物在同一帧或者不同帧的pose。hard negative pair选取的是两个人物bbox有overlap下的pose。构建的训练集具体情况如下：

-	Train	Validation
Positive Pairs	56908	9731
Hard Negative Pairs	25064	7020
Other Negative Pairs	241450	91228

Table 1. Pose pairs collected from PoseTrack'18 dataset.

Ablation Study

Offline vs. Online

离线方法是对每个candidate的每一帧都做detection和pose estimation，然后利用一个flow-based tracker去做追踪(基于文章pose flow)，原理就是把不同帧的pose分配给同一个人。
在线方法只在关键帧做detection。需要注意的是在线方法key-frame的时候只用sptail consistency做data association。

-	**Estimation (mAP)**			**Tracking (MOTA)**		
Method	Wri	Ankl	Total	Wri	Ankl	Total
Offline-CPN101	72.6	68.9	**76.4**	56.1	55.3	**62.4**
Offline-MSRA152	73.6	70.5	**77.3**	58.5	58.5	**64.9**
Online-DET-CPN101-8F	70.5	68.3	**74.0**	52.4	50.3	**58.1**
Online-DET-CPN101-5F	71.7	68.9	**75.1**	53.3	51.0	**59.0**
Online-DET-CPN101-2F	72.4	69.1	**76.0**	54.2	51.5	**60.0**
Online-DET-MSRA152-8F	71.1	69.5	**75.0**	54.6	54.6	**61.0**
Online-DET-MSRA152-5F	72.1	70.4	**76.1**	55.2	55.5	**61.9**
Online-DET-MSRA152-2F	73.3	70.9	**77.2**	56.5	56.6	**63.3**

Table 4. Comparison of offline and online pose tracking results with various keyframe intervals on PoseTrack'18 validation set.

离线精度显然高于在线，但是作者认为gap不算大。

GCN vs. Spatial Consistency(SC)

Method	Detect	Keyframe	MOTA	
			CPN101	MSRA152
SC	GT	8F	68.2	72.0
SC+GCN			68.9	72.6
SC		5F	68.7	73.0
SC+GCN			69.2	73.5
SC		2F	72.0	76.7
SC+GCN			73.5	78.0
SC	DET	8F	58.1	61.0
SC+GCN			59.0	62.1
SC		5F	59.0	61.9
SC+GCN			60.1	63.1
SC		2F	60.0	63.3
SC+GCN			61.3	64.6

Table 5. Performance comparison of LightTrack with GCN and SC on PoseTrack'18 validation set.

用SGCN能涨差不多1个点左右，有时候会出现人物的动作很相似的情况，所以优先采用spatial consistency，如下图所示：

Figure 6. In some situations, different people indeed have very similar poses. Therefore, spatial consistency is considered first.

GCN vs. Euclidean Distance(ED)

- ED 95% acc on validation pairs
- GCN 92% acc

结果对比

这个最终结果的对比选取的是PoseTracl'17 test set。
作者的pose estimator训练数据用的是PoseTrack'17 training set和COCO train+val set。

	Method	Wrist-AP	Ankles-AP	mAP	MOTA	fps
	Posetrack 2017 Test Set					
Offline	PoseTrack, CVPR'18 [3]	54.3	49.2	59.4	48.4	-
Offline	BUTD, ICCV'17 [19]	52.9	42.6	59.1	50.6	-
Offline	Detect-and-track, CVPR'18 [12]	-	-	59.6	51.8	-
Offline	Flowtrack-152, ECCV'18 [36]	71.5	65.7	74.6	57.8	-
Offline	HRNet, CVPR'19[33]	72.0	67.0	74.9	57.9	-
Offline	**Ours-CPN101 (offline)**	**68.0 / 59.7**	**62.6 / 56.3**	**70.7 / 63.9**	**55.1**	-
Offline	**Ours-MSRA152 (offline)**	**68.9 / 61.8**	**63.2 / 58.4**	**71.5 / 65.7**	**57.0**	-
Offline	**Ours-manifold (offline)**	**- / 64.6**	**- / 58.4**	**- / 66.7**	**58.0**	-
Online	PoseFlow, BMVC'18 [37]	59.0	57.9	63.0	51.0	10*
Online	JointFlow, Arxiv'18 [10]	53.1	50.4	63.3	53.1	0.2
Online	**Ours-CPN101-LightTrack-3F**	**61.2**	**57.6**	**63.8**	**52.3**	**47* / 0.8**
Online	**Ours-MSRA152-LightTrack-3F**	**63.8**	**59.1**	**66.5**	**55.1**	**48* / 0.7**
	Posetrack 2018 Validation Set					
	Ours-CPN101 (offline)	72.6 / 63.9	68.9 / 62.6	76.4 / 69.7	62.4	-
	Ours-MSRA152 (offline)	73.6 / 65.6	70.5 / 64.9	77.3 / 71.2	64.9	-
	Ours-YoloMD-LightTrack-2F	62.9 / 56.2	57.8 / 53.3	70.4 / 66.0	55.7	59* / 1.9
	Ours-CPN101-LightTrack-2F	72.4 / 66.3	69.1 / 64.2	76.0 / 70.3	61.3	47* / 0.8
	Ours-MSRA152-LightTrack-2F	73.3 / 66.4	70.9 / 66.1	77.2 / 72.4	64.6	48* / 0.7

Table 6. Performance comparison on Posetrack dataset. The last column shows the speed in frames per second (* means excluding pose inference time). For our online methods, mAP are provided after keypoints dropping. For our offline methods, mAP are provided both before (left) and after (right) keypoints dropping.

速度

Telsa P40 GPU:

- 2.9ms pose matching
- 对整个PoseTrack'19 vadation set (74 vids, 8857 frames)，对于在线算法CPN101-LightTack，整个系统的平均时间为.76fps。其中不包括pose estimation的话有47.11fps。
- 总共追踪57298个人，平均每帧跟踪6.54个人。CPN101的速度为140ms。 109ms的姿态检测和31ms的预处理和后处理。
- 作者认为用light的SPE和detetor可以加快模型的速度。

讨论和结论

作者认为这个pipline也可以做SPT，只要在第一帧选择感兴趣的目标就行了，可以有效减小运算量。然后key-frame的时候选择目标位置。
但是显然还是不合适的，因为detection检测的是所有目标，看起来并不是很efficient。