Automatic Adaptation of Person Association for Multiview Tracking in Group Activities

Minh Vo[1], Ersin Yumer[2], Kalyan Sunkavalli[3], Sunil Hadap[3], Yaser Sheikh[1], and Srinivasa G. Narasimhan[1]

[1]Carnegie Mellon University, [2]Argo AI, [3]Adobe Research

Abstract. Reliable markerless motion tracking of multiple people participating in complex group activity from multiple handheld cameras is challenging due to frequent occlusions, strong viewpoint and appearance variations, and asynchronous video streams. The key to solving this problem is to reliably associate the same person across distant viewpoint and temporal instances. In this work, we combine motion tracking, mutual exclusion constraints, and multiview geometry in a multitask learning framework to automatically adapt a generic person appearance descriptor to the domain videos. Tracking is formulated as a spatiotemporally constrained clustering using the adapted person descriptor. Physical human constraints are exploited to reconstruct accurate and consistent 3D skeletons for every person across the entire sequence. We show significant improvement in association accuracy (up to 18%) in events with up to 60 people and 3D human skeleton reconstruction (5 to 10 times) over the baseline for events captured "in the wild".

Keywords: Descriptor adaptation, people association, motion tracking

1 Introduction

With the rapid proliferation of consumer cameras, events are increasingly being recorded from multiple views, such as surprise parties, group games with headmounted "action cams", and sports events. The challenges in tracking and reconstructing such events include: (a) large scale variation (close-up and distant shots), (b) people going in and out of the fields of view many times, (c) strong view point variation, frequent occlusions and complex actions, (d) clothing with virtually no features or clothing that all look alike (school uniforms or sports gear), and (e) lack of calibration and synchronization between cameras. As a result, tracking methods (both single [1–3] and multi-view [4–6]) that rely on motion continuity produces short tracklets. And tracking-by-association methods relying on pretrained descriptors [7, 8] fail to bridge the domain differences between training data captured in (semi-)controlled environments and event videos captured in open settings.

We present a novel approach that integrates tracking-by-continuity and tracking-by-association to overcome both their limitations. We show that even a state-of-art pretrained person appearance descriptor is not sufficient to discriminate

(a) Chasing (b) Tagging (c) Halloween

Fig. 1: Our testing scenes. Chasing has 14 people, 6 of them with camouflage and 3 others with dark clothing, and cannot be distinguished without strong attention to detail. Tagging has 14 people with feature-less clothing making feature tracking hard. Halloween is from an actual suprise birthday during the Halloween party with 60 people and suffers from significant motion blur. There are no constraints on the scene and camera behavior for any of these sequences.

different people over long durations and across multiple views. We bridge the domain gap by refining the pretrained descriptor to the event videos of interest without any manual interventions (like labeling). This self-supervision is achieved by exploiting three specific sources of information in the target domain: (a) short tracklets from tracking-by-continuity methods, (b) multi-view geometry constraints, and (c) mutual exclusion constraints (one person cannot be at two locations at the same time). These constraints allow us to define contrastive and triplet losses [9, 10] on triplets of people images – two of the same person and one of a different one. Even using the most conservative definition of the constraint satisfaction (tiny tracklets, strict thresholds on distance to epipolar lines) allows us to generate millions of triplets for domain adaptation.

While the above domain adaptation stage improves the descriptor discriminability of people with similar appearance, it could also lead to strong semantic bias for people rarely seen in the videos. We address this problem using a multi-task learning objective and jointly optimize the descriptor discrimination on the large labeled corpus of multiple publicly available human re-Identification (reID) datasets and the unlabeled domain videos. A strong person descriptor enables the use of clustering for people tracking. We adopt the clustering framework of Shah and Koltun [11] and enforce soft spatiotemporal constrains from our mined triplets during the construction of the clustering connectivity graph. Since the association is solved globally, there is no tracking drift.

We validate our association accuracy on three highly challenging sequences of complex and highly dynamic group activity: Chasing, Tagging, and Halloween party, captured by at least 14 handheld smartphone and head mounted cameras (see Fig. 1 and Tab. 2 for the scene statistics). Our method shows significant accuracy improvement over the state-of-art pretrained human reID model (18%, 9%, amd 9%, respectively).

These numerical improvements do not tell the full story. To demonstrate the impact of the improvement, we use our association to drive a complete pipeline for 3D human tracking (that exploits physical constraints on human limb lengths and symmetry) to estimate spatially stable and temporally coherent 3D skeleton for each tracked person. Compared to the baseline, our method shows significant improvement (5-10X) in 3D skeleton reconstruction, stability, minimizing

tracking noise. We believe, for the first time, stable and long duration 3D human tracking has been demonstrated in actual chaotic live group events captured in the wild. **Please see supplementary material for reconstructions.**

Contributions: (1) We present an automatic domain adaptation of person appearance descriptor using monocular motion tracking, mutual exclusive constraints, and multiview geometry in a multitask learning framework without any manual annotations. (2) We show that discriminative appearance descriptor enables reliable and accurate tracking via simple clustering. (3) We present a pipeline for accurate and consistent 3D skeleton tracking of multiple people participating in a complex group activity from mobile cameras *"in the wild"*.

2 Related Work

Our work is related to the themes of people reID and multiview motion tracking. People reID focuses on learning appearance descriptors that match people across viewpoints and time. Recent advances in people reID can be attributed to large and high-quality datasets [12–14], and strong end-to-end descriptor learning. Common approaches include verification models [12, 15, 16], classification models [17, 18], or their combinations [19, 20]. Some recent works also consider body part information [21, 22] for fine-grained descriptor learning. We build on these works but show how a generic person descriptor is insufficient for reliable human association on the multiview videos captured in the wild. Instead, we propose an automatic mechanism to adapt the person descriptor to the captured scene without any manual annotations. Thus, our association model is event (scene) specific rather than generic human reID models.

People tracking approaches formulate person association as a global graph optimization problem by exploiting the continuity of object motion; examples include [1, 23, 24] for single view tracking, and [4, 25, 26] for multiview tracking from surveillance cameras. These approaches use relatively simple appearance cues such as the histogram of color, optical flow, or just the overlapping bounding box area [2, 3, 23, 27–30] for monocular settings or 3D occupancy map from multiview systems [5, 31, 32]. These methods mainly focus on reliable short-term tracklets as the targets permanently disappear after a short time. Our algorithm takes those tracklets as inputs and produces their associations. Additionally, whereas existing multiview tracking algorithms require calibrated and mostly stationary cameras [5, 31–33], our method can handle uncalibrated moving cameras and can temporally align multiple videos automatically during the data generation process for domain adaption.

There are also recent efforts to combine the benefits of global optimization for people association with discriminative appearance descriptors with clear improvements over isolated approaches [34–38]. Notably, Yu et al. [8] present an identity-aware multiview tracking algorithm for a *known number of people* that exploits the sparsely available face recognition, mutual exclusion constraints, and the locality information on a 3D ground plane obtained from fixed cameras to solve a L_0 optimization problem. We address a similar problem but in

unconstrained settings with handheld cameras and unknown number of people. Our insight is to learn strong scene-aware person descriptor rather than solving complex optimization problems.

Our application to 3D markerless motion tracking has been studied in both laboratory setups [39–42] and more general settings with less restrictive model assumptions [43], owing to advances in CNN-based body pose detectors [44–46]. Recent methods with sophisticated visibility modeling [47] or learning based regression [48, 49] enabling motion tracking with few or monocular camera while trading off accuracy are actively explored. However, existing methods for motion tracking showcase the results on activity involving 1 or 2 people in restricted setups. In contrast, we target 3D motion tracking of complex group activities of many people (up to 60 people) in unconstrained settings.

3 Scene Aware Human Appearance Descriptor

Our goal is to learn a robust appearance descriptor extractor $u_x = f(x)$ of a person image x that is similar for images of the same person and dissimilar for different people regardless of the viewing direction, pose deformation, and other factors (like illumination) for our domain videos. We start with an extractor $f(x)$, initially trained on a large labeled corpus of multiple publicly available people ReID datasets, and finetune it using the Siamese triplet loss on triplets of images automatically mined from the domain videos. While this finetuning stage improves the descriptor discriminability of people with similar appearance, it could also lead to strong semantic bias for people rarely seen in the videos. We address this problem using a multitask learning objective and jointly optimize the descriptor discriminability on the labeled corpus labeled human ReID datasets and the unlabeled domain videos.

3.1 Person Appearance Descriptor Adaptation

Due to possible discrepancies between the appearances of the training sets and our domain application videos, we finetune the CNN model on each of our test video sequences using the contrastive and triplet loss [9, 10]. The input to our process are triplets of 2 images of the same person and 1 image of a different person. We optimize the CNN such that the distance between query and anchor is small and the distance between query and the negative example is large. Our loss function is defined as:

$$
\begin{aligned}
L_S(u_i, u_i^+, u_i^-) &= \|u_i - u_i^+\|_2^2 + \max\left(0, \|u_i - u_i^-\|_2^2 - m\right) + \max\left(0, \|u_i^+ - u_i^-\|_2^2 - m\right), \\
L_T(u_i, u_i^+, u_i^-) &= \max\left(0, \|u_i - u_i^+\|_2^2 - \|u_i - u_i^-\|_2^2 + m\right), \\
L_{ST}(u_i, u_i^+, u_i^-) &= L_S(u_i, u_i^+, u_i^-) + L_T(u_i, u_i^+, u_i^-),
\end{aligned}
$$

where, $\{u_i, u_i^+, u_i^-\}$ is a triplet of two positive and a negative unit norm descriptor, respectively, and m (set to 2 for all experiments) is the margin parameter

between two distances. Our total loss function for finetuning is defined as:

$$E_{ST} = \min_{f} \sum_{i=1}^{N_d} L_{ST}(u_i, u_i^+, u_i^-),$$

where, N_d is the number of triplets in the domain videos. We optimize the model using SGD. No hard-negative mining is used due to possibly erroneous labeling.

Automatic Triplet Generation

Single-view triplets: For every video, we first apply CPM [44] to detect all the people and their corresponding anatomical keypoints. Given these detections, we can easily generate negative pairs by exploiting mutual exclusive constraints, i.e. the same person cannot appear twice in the same image. In addition, we can create positive pairs by using short-term motion tracking. We create motion tracklets by combining three trackers: bidirectional Lucas-Kanade tracking of the keypoints, bidirectional Lucas Kanade tracking of the Difference of Gaussian features found within the detected person bounding box, and person descriptor matching between consecutive frames. The tracklet is split whenever any of the trackers disagree. We also monitor the smoothness of the keypoints and split the tracklet whenever the instantaneous 2D velocity is 3 times greater than its current average value. More sophisticated approaches such as [23, 24] can also be used for better tracklet generation. Images corresponding to the same motion tracklet constitute positive pairs for our finetuning.

Multi-view triplets: We enrich the training samples to generate positive pairs across views by using multiview geometry – pairs of detections corresponding to a single person in 3D space should satisfy epipolar constraints. Since our videos are captured in the wild, they are unlikely to be synchronized. Thus, we must first estimate the temporal alignment between cameras to use multiview geometry constraints. Assuming known camera framerate and start time from the video metadata, which aligns the videos up to a few seconds, we linearly search for the temporal offset with the highest number of inliers satisfying the fundamental matrix. A bi-product of the temporal alignment process is the corresponding tracklets across views, which form our positive pairs.

More specifically, let $\mathbf{k}_i^n(t) = \{k_i^{n_{t,1}}, ..., k_i^{n_{t,18}}\}$ be the set of anatomical keypoints of the people detection n at frame t of camera i, and $\mathbf{T}_i^l = \{n_0, .., n_F\}$ be a tracklet l containing the images of the same person for F frames. Let $M_c = (\mathbf{T}_i^l, \mathbf{T}_j^k)$ be the candidate tracklet pair c of the same person, computed by examining the median of the cosine similarity score of between all pairs of descriptors [1] within the tracklets, for camera pair (i, j) and $\mathbf{M}_{i,j}$ be all putative matched tracklets for camera pair (i, j). We set the similarity threshold to 0.5 and add those candidate matches to the hypothesis pool until their ratio-test

[1] At this stage, the descriptors are extracted using a pretrained CNN model. Please refer to the supplementary material for more details about this model.

threshold drops below 0.7. We use RANSAC with the point-to-line (epipolar line) distance as the scoring criteria to try all possible time offsets within the window of $[-2W, 2W]$ frames to detect the hypothesis with the highest number of geometrically consistent matched tracklets:

$$I \leftarrow \mathbf{RANSAC}_{M_c \in \mathbf{M}_{i,j}} \sum_{w=-W}^{W} \sum_{t=1}^{F} \sum_{\substack{n=1 \\ n \in \mathbf{T}_i^l(t) \\ m = \mathbf{T}_j^k(t+w) \\ (\mathbf{T}_i^l, \mathbf{T}_j^k) \in M_c}}^{N_i(t)} \sum_{p=1}^{18} d(k_i^{n,p}, k_j^{m,p}, \mathbf{F}_{i,j}(t)),$$

where, $N_i(t)$ is the number of people detected in camera i at frame t, I is the number of inliers, and $d(x_1, x_2, \mathbf{F}_{i,j}(t))$ is the bidirectional point-to-line distance characterized by the fundamental matrix $\mathbf{F}_{i,j}(t)$ between the camera pair. $\mathbf{F}_{i,j}(t)$ can either be estimated by calibrating the cameras with respect to the scene background or explicitly searched for using the body keypoints during the time alignment process. We prune erroneous matches by enforcing cycle-consistency within any triplet of cameras with overlapping field of view. We set W to twice the camera framerate and use the video start time to limit the search.

3.2 Multitask Person Descriptor Learning

While finetuning the person appearance descriptor exclusively on the test domain could potentially improve discrimination of similar looking people, using it alone may result in semantic drift. The learned descriptor has a strong bias toward frequently observed people, and the descriptor of different people who are rarely observed together from a single camera cannot be forced to be different.

Thus, we jointly learn the person descriptor from both the large scale labeled human identity training data and the scene specific videos. Since the model must predict the identity of the person from the labeled dataset, it is expected to output discriminative descriptors for rarely seen people in the domain videos. On the other hand, since we finetune the model on the domain videos, it should also discriminate people in those sequences better than training solely on other datasets. Mathematically, our multitask loss function is defined as:

$$E_D = \min_f (1 - \alpha) E_{SM} + \alpha E_{ST},$$

where α is the scalar balancing the contribution of two learning tasks. E_{SM} is the standard classification loss:

$$E_{SM} = \operatorname*{argmin}_f \sum_i^{N_s} L_{SM}\big(g(f(x_i)), y_i\big),$$

where, N_s is the number of training examples in the labeled corpus datasets, g is a linear function mapping the person appearance descriptor, $f(\cdot)$, to a vector of the dimension of the number of people in the training corpus, and L_{SM} is the softmax loss penalizing wrong prediction of the people ID label. We set α equal to 0.5 for all experiments.

4 Multiview Tracking via Constrained Clustering

Given the person descriptor, we group detections of the same person across all space-time instances. We rely on the clustering framework of Shah and Koltun [11] but explicitly enforce soft constraints from motion tracklets, mutual exclusive constraints, and geometry matching to link detections. This clustering is formulated as the optimization problem:

$$C = \min_{\mathbf{m}} \sum_{i=1}^{N} \|u_i - m_i\|_2^2 + \lambda \sum_{(p,q)\in Q} w_{p,q}\rho(\|m_p - m_q\|_2),$$

where, N is the number of people detectors, Q is the set of edges in a graph connecting data points u_i, $\mathbf{m} = \{m_1, .., m_N\}$ are the representative of the input descriptors $\mathbf{u}$, λ is a balancing scalar, and ρ is the German-McClure estimator. $w_{p,q} = \frac{\sum_i^N N_i}{N\sqrt{N_p N_q}}$, where N_i is the number of edges connecting x_i in Q, balances the strength of the connection (p, q). Depending on the discrimination of u, the correct number of cluster can be automatically determined during the optimization process [11].

In our settings, we first compute the similarity between tracklet descriptors by taking the median of all possible pairs within the two tracklets to construct the mutual k-NN graph [50]. The number of nearest neighbors is chosen such that the distance between different tracklets is 2 times larger than the median of the tracklet self-similarity score. All detectors belonging to the same tracklet are connected with detectors of their k mutually nearest tracklets. We then add/prune connections that satisfy/violate the multiview triplets mined in Section 3.1. All positive pairs of the triplets are connected, and all negative pairs are disconnected. Finally, we remove the connectivity for detections with no overlapping camera viewing frustums.

5 Application: Human-aware 3D Tracking

To show the benefit of our scene aware descriptor, we build a pipeline for markerless motion tracking of complex group activity from handheld cameras. We first cluster the descriptors from all camera to obtain person tracking information. We then exploit human physical constraints on limb length and symmetry to estimate spatially stable and temporally coherent 3D skeleton for each person.

For each person (cluster), we wish to estimate a temporally and physically consistent human skeleton model for the entire sequence. This is achieved by minimizing an energy function that combines an image observation cost, motion coherence, and a prior on human shape:

$$E(\mathbf{K}, \overline{\mathbf{L}}) = E_I(\mathbf{K}) + E_L(\mathbf{K}, \overline{\mathbf{L}}) + E_S(\mathbf{K}) + E_M(\mathbf{K}), \quad (1)$$

where, $\mathbf{K}$ is the 3D location of the anatomical keypoints over the entire sequence, $\overline{\mathbf{L}}$ is the set of mean limb length for each person. The image evidence cost E_I

encourages the image reprojection of the set of keypoints 3D position to be close to the detected 2D keypoints. The human constant limb length cost E_L minimizes the variations of the human limb length over the entire sequence. The left-right symmetric cost E_S penalizes large bone length differences between the left and right side of the person. The motion coherency cost E_M prefers trajectory of constant velocity [51]. The formulation for each of these terms are given in Table 1. We weight these costs equally.

$E_I(\mathbf{K})$	$\sum_{c=1}^{C}\sum_{t=1}^{F}\sum_{n=1}^{N}\sum_{p=1}^{18}\rho\left(V_c^{np}(t)\frac{\pi_c(K^{np},t)-k_c^{np}(t)}{\sigma_I}\right)$
$E_L(\mathbf{K},\overline{\mathbf{L}})$	$\sum_{t=1}^{F}\sum_{n=1}^{N}\sum_{q\in Q}\left(\frac{\overline{L}^{nq}-L_c^{nq}(t)}{\sigma_L}\right)^2$
$E_S(\mathbf{K})$	$\sum_{t=1}^{F}\sum_{n=1}^{N}\sum_{(l,r)\in S}\left(\frac{L_c^{nl}(t)-L_c^{nr}(t)}{\sigma_S}\right)^2$
$E_M(\mathbf{K})$	$\sum_{n=1}^{N}\sum_{p=1}^{18}\sum_{i=1}^{F-1}\left(\frac{K^{np}(i+1)-K^{np}(i}{\sigma_M^p\Delta(i+1,i)}\right)^2$

C: number of cameras
F: number of frames
N: number of tracked people
$\pi_c(K^p,t)$: projection matrix
$V_c^{np}(t)$: visibility indicator
$L^{nq}(t)$: 3D distance between two points
Q: keypoint connectivity set
S: corresponding left and right limb set
Δ: absolute time differences
σ_I: variation in 2D detection
σ_L: variation in bone length
σ_M^p: variation in 3D speed

Table 1: 3D human tracking cost functions.

We initialize $\mathbf{K},\overline{\mathbf{L}}$ by per-frame RANSAC triangulation of the corresponding person obtained from the clustering and minimize Equation 1 using Ceres solver [52].

6 Experimental Results

Scene	Chasing [C]	Tagging [T]	Halloween [H]
# cameras	5 head-mounted + 12 hand-held	14 hand-held	14 hand-held
Video stats.	1920×1080, 60fps, 30s	1920×1080, 60fps, 60s	3840×2160, 30fps, 120s
# people	14	14	60
Tracklet noise	2%	11%	3%

Table 2: Statistics of the testing scenes. Tracklet noise is the percentage of tracklets with at least two people grouped into a single track.

The proposed method is validated on the three sequences: Chasing [C], Tagging [T], and Halloween [H]. In [T], the camera holders are mostly static and appear in low resolution which does not provide enough appearance variation for strong descriptor learning. It also has many noisy single-view tracklets with different people grouped together due to the lack of texture on the clothing and frequent inter-occlusion. There were no constraints on the camera motion or the scene behavior for any sequence. Refer to Tab. 2 for the statistics of the scenes. We manually annotate the people ID in all sequences for quantitative evaluations.

To perform 3D tracking, we calibrate the cameras using Colmap [53] for [C] and [T]. Due to human motion which frequently occludes the background and strong motion blur, we fail to estimate the camera pose at every frame for [H].

6.1 Analysis of the Descriptor Adaptation

Initially, we pretrain the generic person descriptor on a large corpus that consists of 15 publicly available reID datasets. The combined dataset provides strong diversity in viewpoints, e.g., single camera tracking vs., multiview surveillance

system, appearances, e.g., campus, shopping mall, streets, or laboratory studio, and image resolution. We augment the image with the heatmaps of the body pose provided by CPM [44] to train a pose insensitive person descriptor extractor. This model produces state-of-art descriptor matching on multiple standard benchmarks. Please refer to the supplementary material for details of this model.

Fig. 2 shows 10-NN cross-view matching of images of several people with similar appearance or motion blur for all sequences and their cosine similarity score using the pretrained model and our multitask descriptor learning. The pretrained model retrieves multiple incorrect matches. Our method is notably more accurate. Also, the similarity score often has clear transition between correct and incorrect retrievals. Fig. 3 shows a comparison of the 2D t-SNE embedding [54] between the descriptors using the pretrained model and our multitask learning approach. Our descriptors cleanly group images of the same person together.

We quantify the association accuracy in Fig. 4. For the all sequences, scene specific adaptation of the pre-trained descriptor improves the discrimination of frequently visible actors: 94% vs. 68% 1-NN classification accuracy for [C] and 90% vs. 75% for [T]. However, the discrimination of descriptor for the camera holders decreases: 56% vs. 85% for [C] and 35% vs. 42% for [T]. Our multitask descriptor learning, combining the strength of the classification and metric learning loss, improves both these case (92%/95% for actors/holders on [C] and 89%/61% for [T]) and has an overall baseline improvement of 17% [C], 9% for [T][2], and 9% for [H]. Many false matches due to the confusing appearance descriptor extracted from the generic CNN model are suppressed.

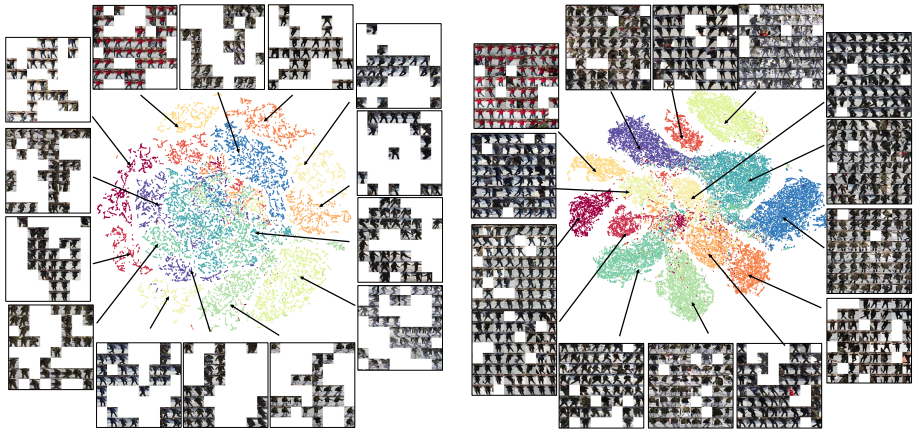
Fig. 5 shows our analysis of the number of cameras, the tracklet noise, and the training videos length on 1-NN matching accuracy. We notice that multiview constraints are more helpful than temporal constraints as there are small improvements compared to the pretrained model P when 1 or 2 cameras are used (mostly corresponding to the small baseline cameras hold by one person). The improvement is saturated when more than 6 cameras are used. Regarding tracklet noise, our algorithm can improve the baseline if the noise percentage is less than 4%. High noise leads to fewer, and potentially incorrect, multiview tracklets from pairwise matches and leads to slightly inferior accuracy compared to P. Lastly, even finetuning on $1/6th$ of the sequences leads to a notable improvement over P and performance converges after $2/6th$ of the sequence is used; this indicates that our method could be used on a smaller training set (e.g., first 15 minutes of a game) and applied to the rest.

6.2 Analysis of Descriptor Clustering

Since each video could contain tens of thousands of people detections, clustering all detectors for all videos jointly could be computationally costly. We adaptively sample the people detector according to their 2D proximity with other detectors and the speed of the detector within each tracklet. All close-by detectors are sampled. Detectors that can be linearly interpolated by others within the same tracklet are ignored. Detectors with less than 9 keypoints detected are also

[2] The results for [T] was obtained with cleaned tracklets.

Fig. 2: 10-NN cross-view matching of the several people with confusing appearance and their cosine similarity score using the pretrained model and our multitask descriptor learning (MTL). Green denotes the query and red denotes incorrect matches. We label the query in green and wrong association in red. Our method retrieves more positive matches and provides easy-to-separate similarity score. All top three neighbors are of the same person.



(a) Pretrained (b) MTL

Fig. 3: t-SNE visualization of the person descriptor extracted using a pretrained model and our multitask learning (MTL) for sequence [C]. Except for images of the same tracklet within a single view, the pretrained descriptors are scatter. Our descriptor groups images of the same person from all views and time instances into cleanly separated clusters. See Fig. 4 for extra quantitative evidences.

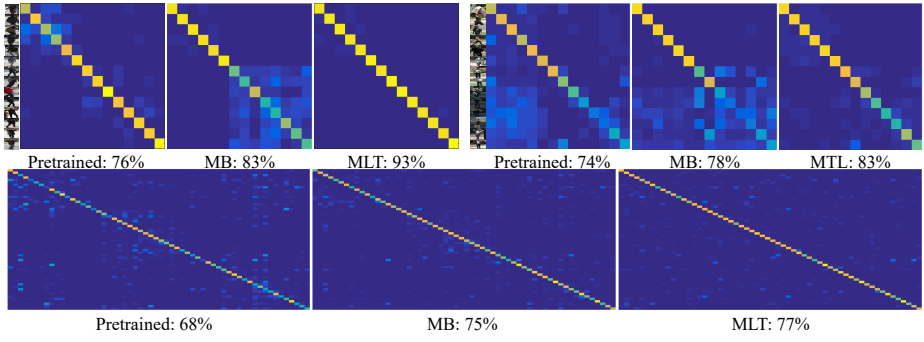


Fig. 4: The confusion matrix of the top-1 matches for the all sequences ([C] top left, [T] top right, [H] bottom) at different stages: pretrained model, multiview bootstrapping (MB), and multitask learning (MTL). There are consistent improvements in accuracy as more sophisticated stage is executed.

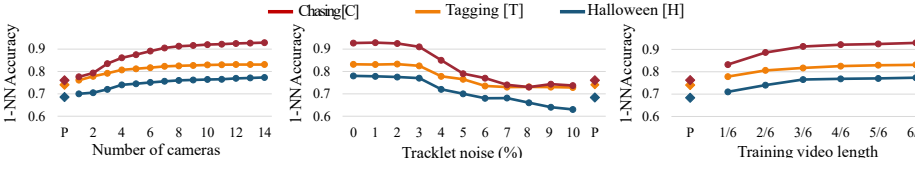


Fig. 5: 1-NN matching accuracy analysis of the proposed method for different number of cameras, percentage of tracklet noise (two or more people grouped in 1 tracklet), and fraction of domain data required for generalization. P denotes the pretrained model. Please refer to the text for the details.

Fig. 6: The 2D projection of the keypoints to all views corresponds well to the expected person anatomical keypoints and tracks people even through occlusions. **Please see supplementary material for the result of [T].**

ignored as they are not very reliably grouped which may hurt subsequent 3D reconstruction. These detectors usually correspond to partially occluded people.

Tab. 3 quantifies the performance of different descriptor learning algorithms by the number of clusters automatically determined by the algorithm, the Adjusted Rand Index (ARI)[3], and cluster accuracy for all detected people in both sequences. In [T], the algorithm discovers many clusters of the pedestrians who are not participating in the group activity.

	Chasing [C]			Tagging [T]			Halloween[H]		
	#clusters	ARI	Accuracy	#clusters	ARI	Accuracy	#clusters	ARI	Accuracy
Pretrained	21	.88	90.1%	66	.85	86.8%	86	.77	79.5%
MTL	16	.97	98.3%	45	.94	95.6%	71	.85	88.1%

Table 3: Analysis of the supervised clustering by the number of clusters, ARI measure and clustering accuracy. Although all methods detected clusters than needed, they correspond to the small cluster of pedestrians who do not participate in the activity (often seen in [T]) or not fully visible bodies due to occlusion. Clustering our multitask learning (MTL) descriptors achieves near perfect clustering accuracy (98.3% for [C] and 95.6% for [T])))

6.3 Comparison with Previous Methods

Direct comparison with previous methods is not straight forward. We focus on long term tracking of group activities in the wild from multiple uncalibrated and unsynchronized handheld cameras. In contrast, existing reID datasets [12–14]

[3] The ARI is a measure of the similarity between two clusters with different labeling systems and is widely used in statistics [55].

r	[33]	Ours: No tracklets	Ours: Full
0.3	67.0%	71.6%	71.9%
0.5	74.1%	75.8%	76.2%

Table 4: MODA comparison for different radius threshold r on ETHZ.

	Chasing		Tagging	
	Per-frame	Human aware	Per-frame	Human aware
Length Dev. (cm)	7.9	1.5	13.4	1.4
Symmetry Dev. (cm)	9.0	1.2	10.1	1.3

Table 5: Comparison between per-frame 3D skeleton reconstruction using ground truth association and human aware reconstruction. Temporal integration and the physical body constrains improve the 3D skeleton stability by 5 to 10 times.

do not have overlapping cameras capturing the same event, which invalidates our multiview constraint. Prominent single view tracking methods [34, 37, 38] and datasets [56] focus on short term tracklet generation rather than long term tracking. The most similar methods to ours are multiview tracking approaches, though they usually require fixed, calibrated, and synchronized cameras [5, 31–33] and assume non-recurrent behavior of pedestrians in their presented datasets. Nevertheless, we show a comparison of the Multiple Object Detection Accuracy (MODA) with the state-of-the-art multiview tracking method of Baqué et al. [33] on their ETHZ dataset in Tab. 4. Due to large number of negative samples, our method outperforms [33] without using single-view 2D tracking for triplet generation. The accuracy gain of our full method is modest because frequent occlusions and frame sub-sampling prohibit long single-view 2D tracklets.

6.4 Application: 3D Human Skeleton Tracking

As a baseline, we use the ground truth people association to perform a per-frame multiview triangulation along with limb length symmetry constraints link this reconstruction temporally using ground truth person tracking. As shown in Fig. 7, this method does not exploit temporal coherency of the skeleton structure, and fails to obtain smooth and clean human trajectories for [C] and [T]. Our method succeeds despite the strong occlusion and complex motion pattern (see the trajectory evolution). Quantitatively, we show 5 to 10X improvement over the baseline (see Tab. 5). We visualize the reprojection of 3D keypoints to all views for [C] in Fig. 6. The reprojected points are close to the anatomical keypoints. These results demonstrate the applicability of our algorithm to markerless motion capture completely in the wild.

7 Discussion and Conclusion

We have presented a simple and practical pipeline for markerless motion capture of complex group activity from handheld cameras in open settings. This is enabled by our novel, scene-adaptive person descriptor for reliable people association over distant space and time instances. Our descriptor outperforms the baseline by 18% and our 3D skeleton reconstruction is 5-10X more stable than

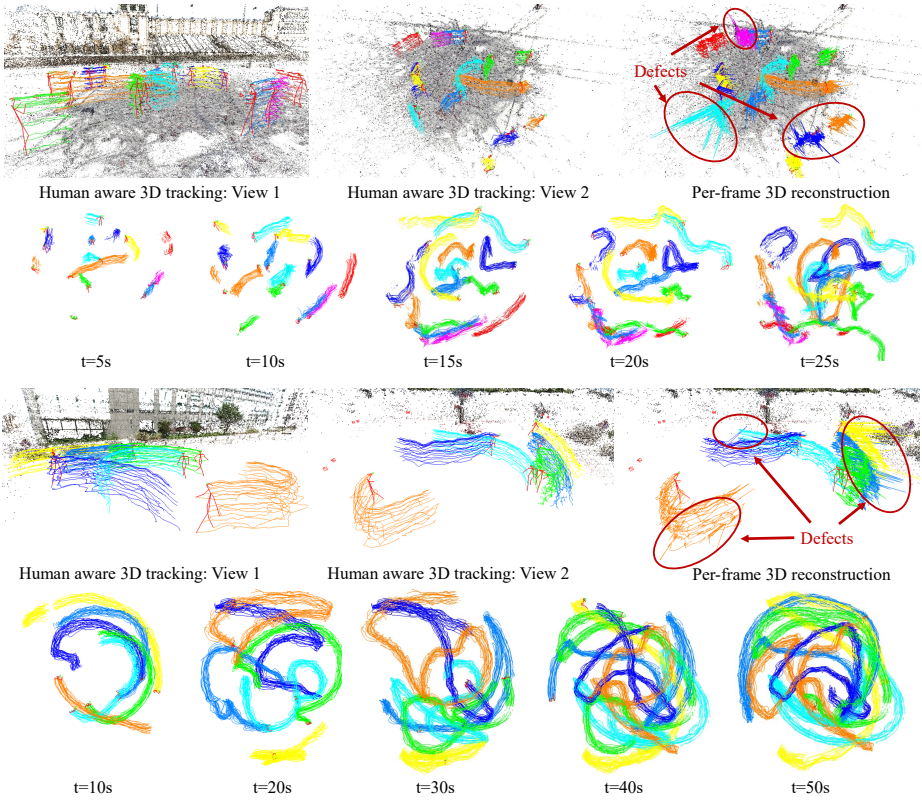


Fig. 7: 3D tracking for [C] and [T] over the entire course of action. The odd rows show the comparison between per-frame 3D reconstruction with ground truth people association and temporally linked for visualization and our algorithm. The even rows show the evolution of the activity using our tracker. Our method succeeds in associating people coming in and out of the camera FoV and gives smooth and clean trajectories despite strong occlusion, similar people appearance, and complex motion pattern. **Please see supplementary material.**

naive reconstruction even with ground truth people correspondences on events captured from handheld cameras in the wild.

Tracklet generation is crucial for descriptor bootstrapping. Noisy tracklets can severely degrade the descriptor discrimination. While more sophisticated algorithms could be used to improve the tracklet generation quality [24, 36], the problem may still remain for scenes with people wearing similar and textureless clothing. One prominent solution is the use of robust estimator for the distance metric loss under the graduated non-convexity framework [11, 57].

Any interesting dynamic event could be overly crowded and people often fully occlude the static background. Since the number of static features observed from any views is significantly smaller than the number of dynamic features, camera localization is very challenging. A feasible solution could use people association and their keypoints to alleviate the need of many static features.

References

1. Zhang, L., Li, Y., Nevatia, R.: Global data association for multi-object tracking using network flows. In: CVPR. (2008)
2. Shitrit, H.B., Berclaz, J., Fleuret, F., Fua, P.: Tracking multiple people under global appearance constraints. In: CVPR. (2011)
3. Choi, W.: Near-online multi-target tracking with aggregated local flow descriptor. In: ICCV. (2015)
4. Berclaz, J., Fleuret, F., Turetken, E., Fua, P.: Multiple object tracking using k-shortest paths optimization. TPAMI (2011)
5. Liem, M.C., Gavrila, D.M.: Joint multi-person detection and tracking from overlapping cameras. CVIU (2014)
6. Rozantsev, A., Sinha, S.N., Dey, D., Fua, P.: Flight dynamics-based recovery of a uav trajectory using ground cameras. In: CVPR. (2017)
7. Assari, S.M., Idrees, H., Shah, M.: Human re-identification in crowd videos using personal, social and environmental constraints. In: ECCV. (2016)
8. Yu, S.I., Meng, D., Zuo, W., Hauptmann, A.: The solution path algorithm for identity-aware multi-object tracking. In: CVPR. (2016)
9. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: CVPR. (2005)
10. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: CVPR. (2015)
11. Shah, S.A., Koltun, V.: Robust continuous clustering. Proceedings of the National Academy of Sciences (2017)
12. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: CVPR. (2014)
13. Zheng, L., Bie, Z., Sun, Y., Wang, J., Su, C., Wang, S., Tian, Q.: Mars: A video benchmark for large-scale person re-identification. In: ECCV. (2016)
14. Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C.: Performance measures and a data set for multi-target, multi-camera tracking. In: ECCVW. (2016)
15. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: CVPR. (2015)
16. Cheng, D., Gong, Y., Zhou, S., Wang, J., Zheng, N.: Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In: CVPR. (2016)
17. Wu, S., Chen, Y.C., Li, X., Wu, A.C., You, J.J., Zheng, W.S.: An enhanced deep feature representation for person re-identification. In: WACV. (2016)
18. Xiao, T., Li, H., Ouyang, W., Wang, X.: Learning deep feature representations with domain guided dropout for person re-identification. In: CVPR. (2016)
19. Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: NIPS. (2014)
20. McLaughlin, N., Martinez del Rincon, J., Miller, P.: Recurrent convolutional network for video-based person re-identification. In: CVPR. (2016)
21. Li, D., Chen, X., Zhang, Z., Huang, K.: Learning deep context-aware features over body and latent parts for person re-identification. In: CVPR. (2017)
22. Zhao, H., Tian, M., Sun, S., Shao, J., Yan, J., Yi, S., Wang, X., Tang, X.: Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In: CVPR. (2017)
23. Milan, A., Roth, S., Schindler, K.: Continuous energy minimization for multitarget tracking. TPAMI (2014)

24. Dehghan, A., Modiri Assari, S., Shah, M.: Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In: CVPR. (2015)
25. Shitrit, H.B., Berclaz, J., Fleuret, F., Fua, P.: Multi-commodity network flow for tracking multiple people. TPAMI (2014)
26. Wang, X., Türetken, E., Fleuret, F., Fua, P.: Tracking interacting objects using intertwined flows. TPAMI (2016)
27. Nguyen, H.T., Smeulders, A.W.: Fast occluded object tracking by a robust appearance filter. TPAMI (2004)
28. Alt, N., Hinterstoisser, S., Navab, N.: Rapid selection of reliable templates for visual tracking. In: CVPR. (2010)
29. Collins, R.T.: Multitarget data association with higher-order motion models. In: CVPR. (2012)
30. Butt, A.A., Collins, R.T.: Multi-target tracking by lagrangian relaxation to min-cost network flow. In: CVPR
31. Fleuret, F., Berclaz, J., Lengagne, R., Fua, P.: Multicamera people tracking with a probabilistic occupancy map. TPAMI (2008)
32. Wu, Z., Thangali, A., Sclaroff, S., Betke, M.: Coupling detection and data association for multiple object tracking. In: CVPR. (2012)
33. Baqué, P., Fleuret, F., Fua, P.: Deep occlusion reasoning for multi-camera multi-target detection. In: ICCV. (2017)
34. Zhang, L., van der Maaten, L.: Structure preserving object tracking. In: CVPR. (2013)
35. Tang, S., Andriluka, M., Milan, A., Schindler, K., Roth, S., Schiele, B.: Learning people detectors for tracking in crowded scenes. In: CVPR. (2013)
36. Dehghan, A., Tian, Y., Torr, P.H., Shah, M.: Target identity-aware network flow for online multiple target tracking. In: CVPR. (2015)
37. Milan, A., Leal-Taixé, L., Schindler, K., Reid, I.: Joint tracking and segmentation of multiple targets. In: CVPR. (2015)
38. Tang, S., Andriluka, M., Andres, B., Schiele, B.: Multi people tracking with lifted multicut and person re-identification. In: CVPR. (2017)
39. Elhayek, A., Stoll, C., Hasler, N., Kim, K.I., Seidel, H.P., Theobalt, C.: Spatio-temporal motion tracking with unsynchronized cameras. In: CVPR. (2012)
40. Liu, Y., Gall, J., Stoll, C., Dai, Q., Seidel, H.P., Theobalt, C.: Markerless motion capture of multiple characters using multiview image segmentation. TPAMI (2013)
41. Stoll, C., Hasler, N., Gall, J., Seidel, H.P., Theobalt, C.: Fast articulated motion tracking using a sums of gaussians body model. In: ICCV. (2011)
42. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: ICCV. (2015)
43. Elhayek, A., de Aguiar, E., Jain, A., Thompson, J., Pishchulin, L., Andriluka, M., Bregler, C., Schiele, B., Theobalt, C.: Marconi: Convnet-based markerless motion capture in outdoor and indoor scenes. TPAMI (2017)
44. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR. (2017)
45. Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P.V., Schiele, B.: Deepcut: Joint subset partition and labeling for multi person pose estimation. In: ECCV. (2016)
46. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: ECVV. (2016)

47. Rhodin, H., Robertini, N., Richardt, C., Seidel, H.P., Theobalt, C.: A versatile scene model with differentiable visibility applied to generative pose estimation. In: ICCV. (2015)
48. Moreno-Noguer, F.: 3d human pose estimation from a single image via distance matrix regression. In: CVPR. (2017)
49. Pavlakos, G., Zhou, X., Derpanis, K.G., Daniilidis, K.: Coarse-to-fine volumetric prediction for single-image 3d human pose. In: CVPR. (2017)
50. Brito, M., Chavez, E., Quiroz, A., Yukich, J.: Connectivity of the mutual k-nearest-neighbor graph in clustering and outlier detection. Statistics & Probability Letters (1997)
51. Vo, M., Narasimhan, S.G., Sheikh, Y.: Spatiotemporal bundle adjustment for dynamic 3d reconstruction. In: CVPR. (2016)
52. Agarwal, S., Mierle, K., Others: Ceres solver. http://ceres-solver.org
53. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR. (2016)
54. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. JMLR (2008)
55. Hubert, L., Arabie, P.: Comparing partitions. Journal of classification (1985)
56. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: MOT16: A benchmark for multi-object tracking. arXiv:1603.00831 [cs] (2016) arXiv: 1603.00831.
57. Zhou, Q.Y., Park, J., Koltun, V.: Fast global registration. In: ECCV. (2016)