Abstract

We present PANDA, the first gigapixel-level human-centric video dataset, for large-scale, long-term, and multi-object visual analysis. The videos in PANDA were captured by a gigapixel camera and cover real-world scenes with both wide field-of-view (~1 km² area) and high-resolution details (~gigapixel-level/frame). The scenes may contain 4k head counts with over 100× scale variation. PANDA provides enriched and hierarchical ground-truth annotations, including 15,974.6k bounding boxes, 111.8k fine-grained attribute labels, 12.7k trajectories, 2.2k groups and 2.9k interactions. We benchmark the human detection and tracking tasks. Due to the vast variance of pedestrian pose, scale, occlusion and trajectory, existing approaches are challenged by both accuracy and efficiency. Given the uniqueness of PANDA with both wide FoV and high resolution, a new task of interaction-aware group detection is introduced. We design a ‘global-to-local zoom-in’ framework, where global trajectories and local interactions are simultaneously encoded, yielding promising results. We believe PANDA will contribute to the community of artificial intelligence and praxeology by understanding human behaviors and interactions in large-scale real-world scenes. PANDA Website: http://www.panda-dataset.com.

1. Introduction

It has been widely recognized that the recent conspicuous success of computer vision techniques, especially the deep learning based ones, rely heavily on large-scale and well-annotated datasets. For example, ImageNet [58] and CIFAR-10/100 [64] are important catalyst for deep convolutional neural networks [33, 41], Pascal VOC [26] and MS COCO [47] for common object detection and segmentation, LFW [34] for face recognition, and Caltech Pedestrians [21] and MOT benchmark [51] for person detection and tracking. Among all these tasks, human-centric visual analysis is fundamentally critical yet challenging. It relates to many sub-tasks, e.g., pedestrian detection, tracking, action recognition, anomaly detection, attribute recognition etc., which attract considerable interests in the last decade [55, 9, 46, 68, 50, 63, 18, 71]. While significant progress has been made, there is a lack of the long-term analysis of crowd activities at large-scale spatio-temporal range with clear local details.

Analyzing the reasons behind, existing datasets [47, 21, 51, 28, 56, 6] suffer an inherent trade-off between the wide FoV and high resolution. Taking the football match as an example, a wide-angle camera may cover the panoramic scene, yet each player faces significant scale variation, suffering very low spatial resolution. Whereas one may use a telephoto lens camera to capture the local details of the particular player, the scope of the contents will be highly restricted to a small FoV. Even though the multiple surveillance camera setup may deliver more information, the requisite of re-identification on scattered video clips highly af-
fects the continuous analysis of the real-world crowd behaviour. All in all, existing human-centric datasets remain constrained by the limited spatial and temporal information provided. The problems of low spatial resolution [44, 54, 28], lack of video information [14, 72, 18, 4], unnatural human appearance and actions [1, 37, 36], and limited scope of activities with short-term annotations [6, 49, 15, 53] lead to inevitable influence for understanding the complicated behaviors and interactions of crowd.

To address aforementioned problems, we propose a new gigaPixel-level human-centric video dataset (PANDA). The videos in PANDA are captured by a gigacamera [70, 7], which is capable of covering a large-scale area full of high resolution details. A representative video example of Marathon is presented in Fig. 1. Such rich information enables PANDA to be a competitive dataset with multi-scale features: (1) globally wide field-of-view where visible area may exceed 1 km², (2) locally high resolution details with gigaPixel-level spatial resolution, (3) temporally long-term crowd activities with 43.7k frames in total, (4) real-world scenes with abundant diversities in human attributes, behavioral patterns, scale, density, occlusion, and interaction. Meanwhile, PANDA is provided with rich and hierarchical ground-truth annotations, with 15,974.6k bounding boxes, 111.8k fine-grained labels, 12.7k trajectories, 2.2k groups and 2.9k interactions in total.

Benefiting from the comprehensive and multiscale information, PANDA facilitates a variety of fundamental yet challenging tasks for image/video based human-centric analysis. We start with the most fundamental detection task. Yet detection on PANDA has to address both the accuracy and efficiency issues. The former one is challenged by the significant scale variation and complex occlusion, while the latter one is highly affected by the gigapixel resolution. Whereafter, the task of tracking is benchmarked. Equipped with the simultaneous large-scale, long-term and multi-object properties, our tracking task is heavily challenged, due to the complex occlusion as well as large-scale and long-term activity existing in real-world scenarios. Moreover, PANDA enables a distinct task of identifying the group relationship of the crowd in the video, termed as interaction-aware group detection. In this task, we propose a novel global-to-local zoom-in framework to reveal the mutual effects between global trajectories and local interactions. Note that these three tasks are inherently correlated. Although detection may bias to local high-resolution detail and tracking may focus on global trajectories, the former promotes the latter significantly. Meanwhile, the spatial-temporal trajectories deduced from detection and tracking serve for group analysis.

In summary, PANDA aims to contribute a standardized dataset to the community, for investigating new algorithms to understand the complicated crowd social behavior in large-scale real-world scenarios. The contributions are summarized as follows.

- We propose a new video dataset with gigapixel-level resolution for human-centric visual analysis. It is the first video dataset endowed with wide FoV and high spatial resolution simultaneously, which is capable of providing sufficient spatial and temporal information from both global scene and local details. Complete and accurate annotations of location, trajectory, attribute, group and intra-group interaction information of crowd are provided.

- We benchmark several state-of-the-art algorithms on PANDA for the fundamental detection and tracking tasks. The results demonstrate that existing methods are heavily challenged from both accuracy and efficiency perspectives, and indicate that it is quite difficult to accurately detect objects in a scene with significant scale variation and track objects that move continuously for a long distance under complex occlusion.

- We introduce a new visual analysis task, termed as interaction-aware group detection, based on the spatial and temporal multi-object interaction. A global-to-local zoom-in framework is proposed to utilize the multi-modal annotations in PANDA, including global trajectories, local face orientations and interactions. Promising results further validate the collaborative effectiveness of global scene and local details provided by PANDA.

By serving the visual tasks related to the long-term analysis of crowd activities at a large-scale spatial-temporal range, we believe PANDA will definitely contribute to the community for understanding the complicated behaviors and interactions of crowd in large-scale real-world scenes, and further boost the intelligence of unmanned systems.

2. Related Work

2.1. Image-based Datasets

The most representative human-centric task on image datasets is human (person or pedestrian) detection. The common object detection datasets, such as PASCAL VOC [26], ImageNet [38], MS COCO [47], Open Images [42] and Objects365 [60] datasets, are not initially designed for human-centric analysis, although they contain human object categories. However, restricted by the narrow FoV, each image only contains limited number of objects, far from enough to describe the crowd behaviour and interaction.

Different terms are used in these datasets, such as “person”, “people”, and “pedestrian”. We uniformly use “human” when there is no ambiguity.
**Pedestrian Detection.** Some pioneer representatives include INRIA [19], ETH [25], TudBrussels [67], and Daimler [23]. Later, Caltech [21], KITTI-D [31], CityPersons [75] and EuroCity Persons [8] datasets with higher quality, larger scale, and more challenging contents are proposed. Most of them are collected via a vehicle-mounted camera through the regular traffic scenario, with limited diversity of pedestrian appearances and occlusions. The latest WiderPerson [76] and CrowdHuman [61] datasets focus on crowd scenes with many pedestrians. Due to the trade-off between spatial resolution and field of view, existing datasets cannot provide sufficient high resolution local details if the scene becomes larger.

**Group Detection.** Starting with the free-standing conversational groups (FCGs) decades ago [24], the subsequent works try to study the interacting persons characterized by mutual scene locations and poses, known as F-formations [40]. Representatives ones include IDIAP Poster [35], Cocktail Party [72], Coffee Break [18] and GDet [4]. In [14], the problem of structure group along with a dataset is proposed, which defines the way people spatially interact with each other. Recently, pedestrian group Re-Identification (G-ReID) benchmarks like DukeMTMC Group [48] and Road Group [48] are proposed to match a group of persons across different camera views. However, these datasets only support position-aware group detection, lack of the important dynamic interactions.

### 2.2. Video-based Datasets

**Pedestrian Tracking.** It locates pedestrians in a series of frames and find the trajectories of them. MOT Challenge benchmarks [43, 51] were launched to establish a standardized evaluation of multiple objects tracking algorithms. The latest MOT19 benchmark [20] consists of 8 new sequences with very crowded challenging scenes. Besides, some datasets are designed for specific applications, e.g., Campus [57] and VisDrone2018 [79], which are drone-platform-based benchmarks. PoseTrack [2] contains joint position annotations for multiple persons in videos. To increase the FoV for long-term tracking, a network of cameras is adopted, leading to the multi-target multi-camera (MTMC) tracking problem. MARS [77], DukeMTMC [56] are representative ones.

On the other hand, to investigate pedestrians in surveillance perspectives, UCY Crowds-by-Example [44], ETH BIWI Walking Pedestrians [54], Town Center [5] and Train Station [78] are proposed for trajectory prediction, abnormal behaviour detection, and pedestrian motion analysis. PETS09 [28] was collected by eight cameras in a campus for person density estimation, people tracking, event recognition, etc. Recently, CUHK [59] and WorldExpo’10 [74] serve for evaluating the performance of crowd segmentation, crowd density, collectiveness, and cohesiveness estimation. However, these datasets are in insufficient of the richness and complexity of the scenes, and can hardly provide high resolution local details, which is critical to further analyze the human interactions in crowd.

**Interaction Analysis.** SALSA [1] contains uninterrupted multi-modal recordings of indoor social events with 18 participants for over 60 minutes. Panoptic Studio [37] uses 480 synchronized VGA cameras to capture social interactions, with 3D body poses annotated. BEHAVE [6], CAVIAR [49], Collective Activity [15] and Volleyball [36] are widely used datasets to evaluate human group activity recognition approaches. VIRAT [53] is a real-world surveillance video dataset containing diverse examples of multiple types of complex visual events. However, for the sake of local details, the group interactions are usually restricted in small scenes or unnatural human behaviors.

### 3. Data Collection and Annotation

#### 3.1. Data Collection and Pre-processing

It is known that single camera based imaging suffers inevitable contradiction between wide FoV and high spatial resolution. The recently developed array-camera-based gigapixel videography techniques significantly boost the feasibility of high performance imaging [7, 70]. By designing the advanced computational algorithms, a number of micro-cameras work simultaneously to generate a seamless gigapixel-level video in realtime. As a result, the sacrifice in either field of view or spatial resolution can be eliminated. We adopt the latest gigacameras [3, 70] to collect the data for PANDA, where the FoV is around 70 degree horizontally, and the video resolution reaches 25k×14k, working in 30Hz. A representative video Marathon in Fig. 1 fully reflects the uniqueness of PANDA with both globally wide FoV and locally high resolution details.

Currently, PANDA is composed by 21 real-world outdoor scenes2, by taking scenario diversity, pedestrians density, trajectory distribution, group activity, etc. into account. In each scene, we collected approximately 2 hours of 30Hz video as the raw data pool. Afterwards, around 3600 frames (approximately two minutes long segments) are extracted. For the images to be annotated, around 30 representative frames per video, 600 in total are selected, covering different crowd distributions and activities.

#### 3.2. Data Annotation

Annotating PANDA images and videos faces the difficulty of full image annotation due to the gigapixel-level resolution. Herein, following the idea of divide-and-merge, the

---

2We are continuously collecting more videos to enrich our dataset. Note that all the data was collected in public areas where photography is officially approved, and it will be published under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License [17].
3.2.1 Image Annotation

PANDA has 600 well annotated images captured from 21 diverse scenes for the multi-object detection task. Among them, PANDA-Crowd subset are composed by 45 images labeled with human head bounding boxes, which are selected from 3 extremely crowded scenes that full of head-counts. The remaining 555 images from 18 real-world daily scenes own 111k pedestrians in total, labeled with head point, head bounding box, visible body bounding box, and the estimated full body bounding box close to the border of the pedestrian. For the crowd that are too far or too dense to be individually distinguished, the glass reflected persons, and the persons with more than 80% occluded area are marked as ‘ignore’ and disabled for benchmarking.

Fig. 2 presents a typical large-scale real-world scene OCT Harbour in PANDA, where the crowd shows a significant diversity in the scale, location, occlusion, activity, interaction, and so on. Beside the fine bounding boxes in (b), each pedestrian is further assigned a fine-grained label showing the detailed attributes in (c). Five categories are used, i.e., walking, standing, sitting, riding, and held in arms (for child), based on the daily postures. Pedestrians whose key parts occluded are marked as ‘unsure’. The ‘riding’ label is further subdivided into bicycle rider, tricycle rider and motorcycle rider. Another detailed attribute is termed as ‘child’ or ‘adult’, distinguished from the appearance and behavior, as shown in (a).

The comparisons with the representative Caltech [21] and CityPersons image datasets [75] are provided quantitatively (Table 1) and statistically (Fig. 3). From Tab. 1, each image of PANDA owns gigapixel-level resolution, which is around 100 times of existing datasets. Although the number of images is much smaller than other datasets, benefiting from the joint high resolution and wide FoV, PANDA has much higher pedestrian density per image than others especially in the extremely crowded PANDA-Crowd, and maintaining the total number of pedestrian in PANDA comparable to Caltech.

Some detailed statistics about image annotation are

<table>
<thead>
<tr>
<th></th>
<th>Caltech</th>
<th>CityPersons</th>
<th>PANDA</th>
<th>PANDA-C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Res</td>
<td>480P</td>
<td>2048×1024</td>
<td>&gt;25k×14k</td>
<td>&gt;25k×14k</td>
</tr>
<tr>
<td>#Im</td>
<td>249.9k</td>
<td>5k</td>
<td>555</td>
<td>45</td>
</tr>
<tr>
<td>#Ps</td>
<td>289.4k</td>
<td>35.0k</td>
<td>111.8k</td>
<td>122.1k</td>
</tr>
<tr>
<td>Den</td>
<td>1.16</td>
<td>7.0</td>
<td>201.4</td>
<td>2,713.8</td>
</tr>
</tbody>
</table>

Table 1. Pedestrian datasets comparison (statistics of CityPersons only contain public available training set). Res is the image resolution, #Im is the total number of images, #Ps is the total number of persons, Den denotes person density (average number of person per image) and PANDA-C is the PANDA-Crowd subset.
3.2.2 Video Annotation

Video annotation pays more attention on the labels revealing the activity/interaction. In addition to the bounding box of each person, we also label the face orientation (quantified into eight bins) and the occlusion ratio (without, partial and heavy). For pedestrians who are completely occluded for a short time, we label a virtual body bounding box and mark it as ‘disappearing’. MOT annotations are available for all the videos in P ANDA except for PANDA-Crowd.

The comparisons with KITTI-T [31] and MOT [20] video datasets are provided quantitatively (Tab. 2) and statistically (Fig. 3). Apparently, P ANDA is competitive with the largest number of frames, tracks and bounding boxes. Moreover, in Fig. 3(e), we show the distribution of tracking duration of different datasets. It demonstrates that the tracking duration in P ANDA is many times longer than KITTI-T and MOT because P ANDA has wider FoV. This property makes PANDA an excellent dataset for large-scale and long-term tracking. Moreover, we also investigate the duration that each person is occluded and summarize the distribution in Fig. 3(f). It shows that more tracks in PANDA suffer from partial or heavy occlusions, in both absolute number and relative portion, making the tracking task more challenging.

For group annotation, the advance of PANDA with wide-FoV global information, high-resolution local details and temporal activities ensures more reliable annotations for group detection. Unlike existing group-based datasets that focus on either the similarity of global trajectories [54] or the stability of local spatial structure [14], we utilize the social signal processing [65] to label the group attributes at the interaction level.

More specifically, with the annotated bounding boxes, we firstly label the group members based on scene characteristics and social signals such as interpersonal distance [22] and interaction [65]. Afterwards, each group is assigned an category label denoting the relationship, such as acquaintance, family, business, etc., as shown in Fig. 2(d). To enrich the features for group identification, we further label the interactions between members within the group.

Table 2. Comparison of multi-object tracking datasets (statistics of KITTI only contain public available training set). Res means video resolution. #V, #F, #T and #B denote the number of video clips, video frames, tracks and bounding boxes respectively. Den means Density (average number of person per frame).

<table>
<thead>
<tr>
<th></th>
<th>KITTI-T</th>
<th>MOT16</th>
<th>MOT19</th>
<th>PANDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Res</td>
<td>1392×512</td>
<td>1080p</td>
<td>1080p</td>
<td>&gt;25k×14k</td>
</tr>
<tr>
<td>#V</td>
<td>20</td>
<td>14</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>#F</td>
<td>19.1k</td>
<td>11.2k</td>
<td>13.4k</td>
<td>43.7k</td>
</tr>
<tr>
<td>#T</td>
<td>204</td>
<td>1.3k</td>
<td>3.9k</td>
<td>12.7k</td>
</tr>
<tr>
<td>#B</td>
<td>13.4k</td>
<td>292.7k</td>
<td>2,259.2k</td>
<td>15,480.5k</td>
</tr>
<tr>
<td>Den</td>
<td>0.7</td>
<td>26.1</td>
<td>168.6</td>
<td>354.6</td>
</tr>
</tbody>
</table>

Since the moving speed of pedestrians is relatively slow and stable, and the posture of pedestrians rarely changes rapidly and dramatically, we label them sparsely on every k frames (k = 6 to 15 based on the scene content) from the perspective of labeling cost. Here we compare the number of bounding boxes after linear interpolation to the original frame rate.

Figure 3. (a) Distribution of person scale (height in pixel). (b) Distribution of the number of person pairs with different occlusion (measured by IoU) threshold per image. (c) Distribution of persons’ pose labels in P ANDA (WK=Walking, SD=Standing, ST=Sitting, RD=Ridding, HA=Held in arms, US=Unsure; The visible ratio is divided into W/O Occ (>0.9), Partial Occ (0.5 - 0.9), and Heavy Occ (<0.5)). (d) Distribution of categories and duration of inter-group interactions in PANDA (PC=Physical contact, BL=Body language, FE=Face expressions, EC=Eye contact, TK=Talking; The duration is divided into Short (<10s), Middle (10s - 30s), and Long (>30s)). (e) Distribution of person tracking duration. (f) Distribution of person occluded time ratio. The comparisons in (a), (b), (e) and (f) are limited to training sets.

In particular, Fig. 3(a) shows the distribution of person scale in pixel of PANDA, Caltech and CityPersons. As we can see, the height of persons in Caltech and CityPersons is mostly between 50px and 300px due to the limited spatial resolution, while PANDA has more balanced distribution from 100px to 600px. The larger scale variation in PANDA necessitates powerful multi-scale detection algorithms. In Fig. 3(b), the pairwise occlusion between persons measured by bounding box IoU of PANDA and CityPersons is given. The fine-grained label statistics for different poses and occlusion conditions are summarized in Fig. 3(c).
including the interaction category (including physical contact, body language, face expressions, eye contact and talking; multi-label annotation) and its begin/end time. The distribution and duration of interactions are shown in Fig. 3(d). The mean duration of interaction is 518 frames (17.3s). To avoid overly subjective or ambiguous cases, three rounds of cross-checking are performed.

### 4. Algorithm Analysis

We consider three human-centric visual analysis tasks on PANDA. The first is pedestrian detection, which biases local visual information. The second is multi-pedestrian tracking. In this task, global visual clues from different regions are taken into consideration. Based on these two well-defined tasks, we introduce the interaction-aware group detection task. In this task, both global trajectories and local interactions between persons are necessary.

#### 4.1. Detection

Pedestrian detection is a fundamental task for human-centric visual analysis. The extremely high resolution of PANDA makes it possible to detect pedestrians from a long distance. However, the significant variance in scale, posture, and occlusion severely degrade the detection performance. In this paper, we benchmarked several state-of-the-art detection algorithms on PANDA.4.

**Evaluation metrics.** For evaluation, we choose $AP_{50}$ and $AR$ as metrics: $AP_{50}$ is the average precision at $IoU=0.50$ and $AR$ is the average recall with $IoU$ ranging in $[0.5, 0.95]$ with a stride of 0.05.

**Baseline detectors.** We choose Faster R-CNN [55], Cascade R-CNN [9] and RetinaNet [46] as our baseline detectors with ResNet101 [33] backbone. The implementation is based on [11]. To train the gigapixel images on our network, we resize the original size image into multiple scales and partition the image into blocks with appropriate size as neural network input. For the objects cut by block borders, we retain them if the preserved area covers 50%. Similarly, for evaluation, we resize the original image into multiple scales and use the sliding window approach to generate proper size blocks for the detector. For a better analysis of detector performance and limitations, we split test results into subsets according to the object size.

**Results.** We train these 3 detectors from the COCO pretrained weights and evaluate them on three tasks: visible body, full body, and head detection. As shown in Tab. 3, Faster R-CNN, Cascade R-CNN and RetinaNet show the difficulty in detecting small objects, resulting in very low precision and recall. We also apply false analysis on visible body using Faster R-CNN, as illustrated in Fig. 4 left. We can observe that the huge amount of false negatives is the main reason for poor recall. According to the results, it seems quite difficult to accurately detect objects in a scene with very large scale variation (most 100× in PANDA) by the single detector based on existing architectures. More advanced optimization strategies and algorithms are highly demanded for the detection task on extra-large images with large object scale variation, such as scale self-adaptive detectors and efficient global-to-local multi-stage detectors.

#### 4.2. Tracking

Pedestrian tracking aims to associate pedestrians at different spatial positions and temporal frames. The superior properties of PANDA make it naturally suitable for long-term tracking. Yet the complex scenes with crowded pedestrian impose various challenges as well.

**Evaluation metrics.** To evaluate the performance of multiple person tracking algorithms, we adopt the metrics of MOTChallenge [43, 51], including MOTA, MOTP.
shows the results of DeepSORT. 

Table 4. Performance of multiple object tracking methods on PANDA. T is tracker, D is detector, DS, DAN and MD denote the DeepSORT [68], DAN [63] and MOTDT [50] trackers, respectively. ↑ denotes higher is better and vice versa.

<table>
<thead>
<tr>
<th>T</th>
<th>D</th>
<th>MOTA↑</th>
<th>MOTP↑</th>
<th>IDF1↑</th>
<th>FAR↓</th>
<th>MT↑</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CR</td>
<td>24.35</td>
<td>76.31</td>
<td>21.39</td>
<td>15.59</td>
<td>661</td>
</tr>
<tr>
<td></td>
<td>RN</td>
<td>16.36</td>
<td>78.0</td>
<td>15.16</td>
<td>4.32</td>
<td>259</td>
</tr>
<tr>
<td>DAN [63]</td>
<td>FR</td>
<td>25.06</td>
<td>74.81</td>
<td>21.85</td>
<td>25.95</td>
<td>826</td>
</tr>
<tr>
<td></td>
<td>CR</td>
<td>24.24</td>
<td>78.55</td>
<td>20.13</td>
<td>12.42</td>
<td>602</td>
</tr>
<tr>
<td></td>
<td>RN</td>
<td>15.57</td>
<td>79.90</td>
<td>13.43</td>
<td>3.33</td>
<td>227</td>
</tr>
<tr>
<td>MD [50]</td>
<td>FR</td>
<td>13.51</td>
<td>78.82</td>
<td>14.92</td>
<td>6.52</td>
<td>257</td>
</tr>
<tr>
<td></td>
<td>CR</td>
<td>13.54</td>
<td>80.25</td>
<td>14.89</td>
<td>4.41</td>
<td>255</td>
</tr>
<tr>
<td></td>
<td>RN</td>
<td>10.77</td>
<td>80.62</td>
<td>11.86</td>
<td>1.90</td>
<td>162</td>
</tr>
</tbody>
</table>

Figure 5. Influence of target properties on tracker’s MOTA. We divided the pedestrian targets into 3 subsets from easy to hard for each property.

Confidence threshold 0.7 for detection results.

We further analyze the influence of different pedestrian properties, including: (a) tracking duration; (b) tracking distance; (c) moving speed; (d) scale (height); (e) scale variation (the standard deviation of height); (f) occlusion. For each property, we divided the pedestrian targets into 3 subsets from easy to hard. Besides, in order to eliminate the influence of detectors, we used the ground-truth bounding boxes as input here. Fig. 5(b)(c) show that the tracking distance and moving speed are the most influential factors to trackers’ performance. In Fig. 5(a), the impact of tracking duration on tracker performance is not obvious because there are many stationary or slow moving people in the scene.

4.3. Group Detection

Group detection aims at identifying groups of people from crowds. Unlike existing datasets that focus on either the similarity of global trajectories [54] or the stability of local spatial structure [14], the advance of PANDA with joint wide-FoV global information, high-resolution local details and temporal activities imposes rich information for group detection.

Furthermore, as indicated by the recent advances on trajectory embedding [30, 16], trajectory prediction [10, 13] and interaction modeling in video recognition [36, 62, 27], these tasks are strongly correlated to the group detection task. For example, modeling group interaction can help improve the trajectory prediction performance [69, 10, 13], while learning a good trajectory embedding is also beneficial for video action recognition [66, 30, 16]. However, none of previous research has investigated how those multi-modal information can be incorporated into the group detection task. Hence, we propose the interaction-aware group detection task, where video data and multi-modal annotations (spatial-temporal human trajectory, face orientation, and human interaction) are provided as input for group detection.
Figure 6. Global-to-local zoom-in framework for interaction-aware group detection. The Global Trajectory, Local Interaction, Zoom In, and Edge Merging modules are associated. Different color vertices and trajectories stand for different human entities. Line thickness represents the edge weights in the graph. (1) Global Trajectory: Trajectories are firstly fed into LSTM encoder with dropout layer to obtain embedding vectors and then construct a graph where the edge weight is L2 distance between embedding vectors. (2) Zoom In: By repeating inference with dropout activated as Stochastic Sampling [38], $E_{global}$ and $E_{uncertainty}$ are obtained from sample mean and variance respectively. (3) Local Interaction: The local interaction videos corresponding to high uncertainty edges (14) are further checked using video interaction classifier (3DConvNet [32]). (4) Edge Merge and Results: Edges are merged using label propagation [73], and cliques remaining in the graph are the group detection results.

**Framework.** We further design a global-to-local zoom-in framework as shown in Fig. 6 to validate the incremental effectiveness of local visual clues to global trajectories. More specifically, human entities and their relationships are represented as vertices and edges respectively in graph $G = (V, E)$. And features from multiple scales and modalities such as the global trajectory, face orientation vector, and local interaction video are used to generate edge set $E_{global}$ and $E_{local}$. Following a global-to-local strategy [52, 29, 45, 12], $E_{global}$ is firstly obtained by calculating L2 distance in feature space for each trajectory embedding vector, which comes from LSTM encoder like common practice [30]. After that, uncertainty-based [38, 39] and random selection policies are adopted to determine the sub-set of edges that need to be further checked using visual clues. Then, video interaction scores among entities are estimated by spatial-temporal ConvNet [32]. The combinations of obtained edge sets, e.g., $E_{global} \cup E_{local}$ or $E_{global}$, are merged using label propagation [73], and the cliques remaining in the graph are the group detection results. Finally, we can estimate the incremental effectiveness with the performance metrics specified in [14] under different combinations.

**Results.** The details about network structure, zoom-in policy, and edge merging strategy are placed in supplemental material S.4.3.1, S.4.3.2, and S.4.3.3 respectively, while experimental results are shown in Tab. 5. Half metrics [14] including precision, recall, and F1 where group member IoU = 0.5 are used for evaluation. The performance is improved significantly by leveraging $E_{local}$ as well as uncertainty estimation, which further validates the effectiveness of local visual clues provided by PANDA.

![Diagram](image.png)

Table 5. Incremental Effectiveness (half metric [14]). The random zoom-in policy randomly selects several local videos to estimate interaction score while the uncertainty-based one selects local videos depending on the uncertainty estimation from Stochastic Dropout Sample [38].

<table>
<thead>
<tr>
<th>Edge Sets</th>
<th>Zoom In</th>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_{global}$</td>
<td>$I$</td>
<td>0.237</td>
<td>0.120</td>
<td>0.160</td>
</tr>
<tr>
<td>$E_{global} \cup E_{local}$</td>
<td>Random</td>
<td>0.244</td>
<td>0.133</td>
<td>0.172</td>
</tr>
<tr>
<td>$E_{global} \cup E_{local}$</td>
<td>Uncertainty</td>
<td>0.293</td>
<td>0.160</td>
<td>0.207</td>
</tr>
</tbody>
</table>

5. Conclusion

In this paper, we introduced a gigapixel-level video dataset (PANDA) for large-scale, long-term, and multi-object human-centric visual analysis. The videos in PANDA are equipped with both wide FoV and high spatial resolution. Rich and hierarchical annotations are provided. We benchmarked several state-of-the-art algorithms for the fundamental human-centric tasks, pedestrian detection and tracking. The results demonstrate that they are heavily challenged for accuracy due to the significant variance of pedestrian pose, scale, occlusion and trajectory, etc., and efficiency due to the large image size and the huge amount of objects in single frame. Besides, we introduced a new task, termed as interaction-aware group detection based on the characteristics of PANDA. We proposed a global-to-local zoom-in framework which combines both global trajectories and local interactions, yielding promising group detection performance. Based on PANDA, we believe the community will develop new effective and efficient algorithms for understanding complicated behaviors and interactions of crowd in large-scale real-world scenes.
References


[17] Creative Commons. Commons attribution-noncommercial-sharealike 4.0 license. https://creativecommons.org/licenses/by-nc-sa/4.0/, 3


Conducting interaction: Patterns of behavior

Alex Kendall and Yarim Gal. What uncertainties do we need in focused encounters

Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla.

Hanbyul Joo, Hao Liu, Lei Tan, Lin Gui, Bart Nabbe, Mostafa S Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Hayley Hung and Ben Kröse. Detecting f-formations as

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can


Stefano Pellegrini, Andreas Ess, Konrad Schindler, and Luc Van Gool. You’ll never walk alone: Modeling social beh-


