


Please cite the Published Version

Qing, L, Li, L, Xu, S, Huang, Y, Liu, M, Jin, R, Liu, B, Niu, T, Wen, H, Wang, Y, Jiang, X and Peng, Y  (2021) Public Life in Public Space (PLPS): a multi-task, multi-group video dataset for public life research. In: 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 11 October 2021 - 17 October 2021, Montreal, Canada.

DOI: <https://doi.org/10.1109/ICCVW54120.2021.00404>

Publisher: IEEE

Version: Accepted Version

Downloaded from: <https://e-space.mmu.ac.uk/635039/>

Usage rights:  In Copyright

Additional Information: © 2021 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Enquiries:

If you have questions about this document, contact openresearch@mmu.ac.uk. Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from <https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines>)

Public Life in Public Space (PLPS): A multi-task, multi-group video dataset for public life research

Linbo Qing¹, Lindong Li¹, Shengyu Xu¹, Yibo Huang¹, Mei Liu¹, Rulong Jin¹, Bo Liu¹, Tong Niu¹,
Hongqian Wen¹, Yuchen Wang¹, Xue Jiang¹, Yonghong Peng²

1: Sichuan University, Chengdu, China; 2: Manchester Metropolitan University, Manchester, UK

qing_lb@scu.edu.cn; {li_lindong, xushengyu, 2018222050141, 2018222050144, jinrulong, 2019222055261,
2019222050153, 2020222050177, wangyuchen98, 2020222050122}@stu.scu.edu.cn; Y.Peng@mmu.ac.uk

Abstract

Human-computer interaction (HCI) is a multidisciplinary field of study focusing on the design of computer technology and, in particular, the interactions between humans and computers. Public space, between the urban buildings, is an open and accessible area to people. Public life, happening in public spaces, is about human activity, human interaction, expression of human feeling in the wild. Affective behavior analysis in the public space is the basic topic of the public life research, which is the key to achieve HCI applications through comprehensively understanding people's feelings, emotions, social behaviors and their correlations in a 'human-centered' and engaging manner. However, it is a challenging task to design a robust HCI system due to the lack of multi-task datasets (including emotion, behavior, social relations, etc), collected under the uncontrolled conditions in real public spaces. In spite that existing separate datasets in computer vision can somehow meet the requirement of public life research, they are neither captured from real public spaces nor for multiple tasks, which cannot comprehensively support the joint research of public life. To tackle this issue, this paper presents a multi-task, multi-group human-oriented video dataset, namely public life in public space (PLPS). Specifically, multi-tasks in terms of activity recognition, emotion recognition and social relation recognition are integrated for each video data. Multi-group and multi-level labels in terms of individuals, groups, video clips are included in the dataset. With PLPS, more sophisticated computer vision model for comprehensive public life research can be facilitated.

1. Introduction

Public life is what people create when they interact with each other in public spaces, including the streets, parks, squares, and city spaces between buildings. It consists of

the daily activities that people naturally take part in when they spend time with each other outside their homes, workplaces, and cars. It is a driver of physical and mental health, social benefits, economic development, culture and identity, safety, sustainable mobility.

1.1. Public Life Research and Affective Behavior

In order to achieve HCI systems, affective behavior analysis has to consider people's feelings, emotions, social behaviors and their correlation comprehensively. Public life research is a people-oriented subject, aiming to observe various aspects in public spaces [8]. In other word, affective behaviors correspond to the human-oriented observations for the public life in public spaces. Therefore, public life research for comprehensive affective behavior analysis should take emotion estimation, human activity recognition and group understanding via social cues in the wild into account.

Besides, each of these studies in the literatures gives an insight to one aspect of public life while it is also of great importance to concentrate on their correlations for comprehensive affective behavior analysis. This is critically necessary to create machines and robots that are able to interact in a 'human-centered' and engaging manner with people, and effectively serving them as their digital assistants. Meanwhile, computer vision tasks must be extended to meet the 'in-the-wild' realistic in public spaces, such as multiple groups and fuzzy appearance.

Therefore, it is required to collect a multi-task and multi-group dataset from real public spaces for public life research. In this dataset, various attributions should be annotated, including age, gender, group, action/activity, emotion and social relations.

1.2. Related Datasets

To the best of our knowledge, none of existing datasets meets all requirements for public life research, i.e., multiple

tasks, multiple groups and the public space scenarios in the wild. As listed in Table 1, we pay more attention to the action/activity, emotion and social relations.

For activity recognition, there are five main datasets. Volleyball [13] and NBA [42] focus on sports activities. CAE [4], extended by CAD [3] and with 6 categories, does not cover most of activities in the public space such as sitting and playing. And videos from CND [2] are the same scene while scene are various in the real public space. In addition, single-group CAD, CAE and CND are against the fact of multiple groups in the public space.

Group activity describes people's behaviors and emotion shows their feelings. Many datasets are collected for individual emotion recognition, most of them only concentrate on facial traits, including CK+ [28], MMI [34], Oulu-CASIA [46], AffectNet [33], AFEW [6], Aff-Wild [16, 44] and Aff-Wild2 [15, 17, 18, 19, 20, 21]. Most of data of CAER [23], Emotic [22] and IEMOCAP [1] either only contain the upper body or were collected in limited surroundings rather than public spaces. GroupWalk [32] is the closest to the real situation of public space while it was collected in only 8 fixed locations. In above datasets, people always cover too much area of the image which causes no interaction between people and public space.

For social relations, it is high-level attribution and most human behaviors are developed in the relationship context [35]. In the early stage, datasets mainly collected facial images for kinship verification, i.e., [38], UB KinFace [40], Family101 [7], KinFace [10], KFW-I [27] and KFW-II [27]. And another IRD [45] also collected facial images to recognize pair's association, e.g., warm. Recently, PISC [24] and PIPA [36] were collected for detailed relationship recognition such as friend. In PISC and PIPA, there are abundant scene types while most have nothing to do with the public space. Except for these image-based datasets, KFW [41], SRIV [29] and ViSR [26] were constructed for video-based SRR. KFW are used for facial kinship, SRIV and ViSR come from movies with rare public space scene and part body. They are not suitable for the public life research.

1.3. Contributions

In summary, these datasets for one task are not enough to support the public life research for affective behavior analysis. And more importantly, single attribution cannot facilitate the comprehensive understanding for the public space and uniform multi-task recognition hardly comes true. To address this problem, we collected a multi-task and multi-group dataset, namely public life in public space (PLPS), which is a high-quality dataset with multi-scale labels in terms of individual, group and video clip. Its main contributions are summarized as:

- 1) First dataset specially for public life research, including 71 clips from various complex scenes in public spaces.

It will contribute to the study of comprehensive affective behavior analysis and support the HCI systems.

- 2) Multi-scale labels are provided for related studies, including age, gender, group, action/activity, emotion and social relations. Especially, we extend activities from single group to multiple groups, emotions from individual to group and social relations from pair to group.
- 3) Baselines for action/activity, emotion and social relations, in particular, two novel graph-based models for multi-group activity recognition and group SRR.

The paper is organized as follows. Section 2 describes how to collect our data and protect the privacy of citizens. Section 3 explains what labels should be annotated, gives labeling process and conducts quality analysis of annotations. Section 4 presents the label distribution and Section 6 shows our baseline methods and demonstrates their effectiveness. Section 7 concludes this paper and discusses how to promote the future work of public life research.

2. Data Acquisition

Original data. The public space category is various such as square, park and street. Even for the same category, features of public spaces range from the whole to the part because of differences in geographical location, cultural habits and so on. To sense people and objects for further studies in kinds of public spaces, we obtain many videos from different places of different cities in China. It is worth noting that most videos are from Chengdu, the capital city of Sichuan province, which benefits to conduct in-depth research for a specific city. In addition, we emphasize the human perspective in public spaces. That is, cities should be viewed and depicted at level of human eyes [8]. In compliance with Chinese laws, therefore, we fixed smart phones on a tripod at a height of 1.75 meters to shoot the videos with the resolution of 1920×1080 and with the frame rate of 25fps. It is worth noting that each video was tagged with when and where it was taken for possible research in the future.

Selected data. In order to study algorithms, the original videos were selected and cut into short clips with a length of 8-12s and about 300 frames. To ensure the diversity of characters and scenes, no more than three clips were selected from each original video. As listed in Table 2, we obtained 71 clips with over 20,000 frames and nearly 150,000 bounding boxes for the construction of our dataset.

Privacy protection. When obtaining the videos, we took it into account and complied with relevant laws and privacy protection policy. Meanwhile, we also concentrate on the general practice about how to release a dataset. For example, WoodScape [43] was released with original data and a license agreement which enforces the users to strictly ad-

	Dataset	Data Type	Public Space or not	Individual (Pair)	Group
Action (Activity)	CAD, CAE	Video	Partial	✓	✓
	CND	Video	Yes	✓	✓
	Volleybal	Video	-	✓	✓
	NBA	Video	-	-	✓
Emotion	CK+, MMI, Oulu-CASIA	Facial image	-	✓	-
	AffectNet	Image	-	✓	-
	Emotic	Image	Partial	✓	-
	Aff-Wild, Aff-Wild2	Facial Video	-	✓	-
	AFEW, CAER, IEMOCAP	Video	-	✓	-
	GroupWalk	Video	Yes	✓	-
Social Relation	UB KinFace, Family101, KinFace, KFW, IRD	Facial image	-	✓	-
	PIPA, PISC	Image	Partial	✓	-
	KFWW, SRIV	Video	-	✓	-
	ViSR	Video	Partial	✓	-
Multi-task	PLPS (Ours)	Video	Yes	✓	✓

Table 1. Summary of existing datasets related with our PLPS dataset.

Sichuan	Guizhou	He’nan	Gansu	Hubei	Yunnan
Chengdu (50)	Liupanshui (6)	Xuchang (4)	Jinchang (3)	Qianjiang (1)	Chuxiong (1)
Leshan (3) Dazhou (1)	Zunyi (2)				

Table 2. Cities and the corresponding number of clips.

here to the General Data Protection Regulation (GDPR). We will take a similar approach as well.

3. Labeling Process

3.1. Label Category

According to the requirement of public life research for affective behavior analysis, we annotated multi-scale labels as listed in Table 3.

Representing human emotions has been a basic topic of the affective behavior analysis, meanwhile human action/activity and social relations are the vital influencing factors. All of them are also the important points that are focused on in the public life research. The 3-D Valance, Arousal and Dominance Space (VAD-Space) [30], the most usual dimensional emotion representation, is annotated for the human emotions. Considering the compatibility with other datasets [24, 36] and the real situation of public spaces, social relation category includes friend, family, couple, professional, commercial and no relation. Similarly, human action/activity involves NA, crossing, waiting, queueing, talking, dancing, sitting, jogging, playing, riding and doing sport. NA refers to the unspecific action, such as the middle action from standing to sitting.

Given the group observation of public life research, group labels are also annotated and label categories are same as the above except for social relations. Since group

with social relations gathers three or more people, couple is not included in group social relations. In addition, we further annotated the number of people, bounding box (trajectory), group division, gender (infant, child, teenager, adult and old people) and age (male, female, unknown), which are also the important information in the public space [8].

3.2. Annotation Step

Before annotation, we instructed each annotator how to determine the label type and gave a detailed document, elaborating the annotation step, explaining the label definition and showing the corresponding sample. To show the scene types and label details, we present the samples of action/activity, emotion and social relations in the real and complex public spaces in Figure 1. As for the definition of various labels, the action/activity and social relations have been elaborated in other tasks [2, 4, 24] while the VAD-Space is relatively abstract, i.e., valance (the pleasantness of a stimulus), arousal (the intensity of emotion provoked by a stimulus) and dominance (the degree of control exerted by a stimulus). Hence, the samples of action/activity and social relations with the description of other task are enough to determine what labels should be annotated. To further illustrate the correspondence between VAD-Space and the real situation in public spaces, we provide annotators with the explanations of different levels in terms of specific scenarios in Table 4. With abstract concept, specific expla-

Attribution	Age (Infant, Child, Teenager, Adult, Old People), Gender (Male, Female, Unknown), Group
Action/Activity	NA, Crossing, Waiting, Queueing, Talking, Dancing, Sitting, Jogging, Playing, Riding, Doing Sport
Emotion	Valance (-2, -1, 0, +1, +2), Arousal (-2, -1, 0, +1, +2), Dominance (-2, -1, 0, +1, +2)
Social Relation	Friend, Family, Couple, Professional, Commercial, No relation

Table 3. Label categories.

nation and corresponding samples, it is easier to annotate high-quality labels. The next annotation procedure is as follows:

First, we require the annotator to watch the whole video so as to know what emotions to expect. Then, the annotator assigns number for each individual, draw their bounding boxes and label their actions/emotions in each frame of one clip by Labellmg. Finally, after completing the annotation of each frame, the related labels of the entire clip are annotated, including the total number of people, gender, age, group division, group activity, group emotion and group social relation. These label information is the same through all frames so they also can be viewed as labels of each frame.

3.3. Annotation Quality Control

To control the quality of our dataset, we take some measures to ensure consistency of all annotators. One of the most important is that we clear the uniform annotation details through the oral report and description document as mentioned in Section 3.2. In addition, pre-annotation is adopted for problem summary and rule modification.

Even if uniform rules are made, there are still understanding deviations for different annotators hence we conduct the “1-2-3” mechanism that one annotates all labels of a clip, then two check them independently and finally the three annotators meet the consistency for final labels based on the rule that the minority is subordinate to the majority.

As shown in Table 5, we calculate the agreement rates of all labels, which demonstrate the high consistency and evaluate the high quality of the dataset.

4. Label Distribution

Action/Activity. In total, we annotated 149,592 individuals with 79,090 groups. As shown on the left in Figure 2, the number of crossing is maximum, which is in accordance with the fact that most people are crossing in the real public space. The numbers of waiting, queueing, talking, and sitting are roughly equal, which happen in most scenes. Both dancing and playing are about 7000 instances. Because public square dancing always happens after dinner, dim light is bad for shooting in this period, which causes the minority instances of dancing. For the playing, the corresponding groups are children hence its number is naturally less. Three of the least actions are jogging, riding, and doing sport. Jogging and doing sport usually happen in

specific scenarios such as parks and sports fields. For the riding, there are two reasons. One is that people in public spaces tend to be crossing rather than riding and the other is only short span during which cyclist appears in our camera, which is not beneficial for the next recognition. However, their numbers are still greater than 4000. As shown on the right in Figure 2, the group activity distribution is similar to the individual action but the ratios of queueing and dancing are smaller. It is because the number of people in the two groups (i.e., queueing and dancing) is much greater than that in other groups.

Emotion. The emotion label distribution reflects the real situation in the public space. As shown in Figure 3 and 4, people tend to control themselves in the real public space hence mid-level’s instances cover a large portion in spite of valance, arousal, or dominance. As for the label distribution of valance levels, it is hard to see a negative person in the real public space and there is only one person with the valance-level -2, which appears in one clip with 299 frames in our dataset. People are likely to be neutral and positive. But there exist some slight differences in the arousal level and dominance level. That is, the samples of low levels are more than ones of high levels. It is because people in the real public space tend to work on actions/activities with low intensity and their dominance with other people and the objects in that space is weak.

SRR. The annotation of social relations is based on group division and each group is annotated with one social relation. Obviously, the special group with one individual does not belong to any social relations but this individual is no relation with the other individuals. Further, any person pairs are also no relation when the two person are in different groups. In addition, it is noticed that the person pair in one group is also likely to be no relation such as the queueing group. As shown in Figure 5, we first count the numbers of single-person groups and multi-person groups. Their ratio is close to 1:1, which demonstrates that half of persons in the public space are with their friends, family, lovers, etc. From the sub-pie in Figure 5, it can be observed that the instances of professional and commercial are less than the other social relations, which is because the two relations happen indoor. This slight imbalance is not enough to influence the study of algorithm.

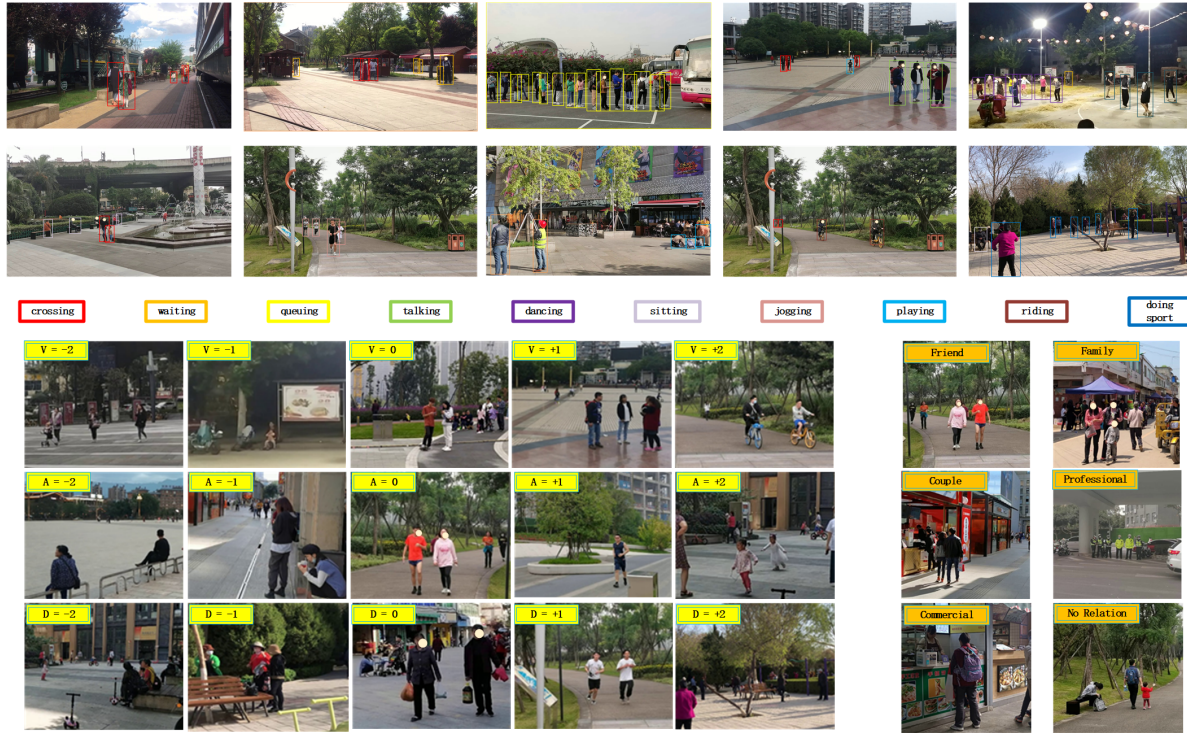


Figure 1. Samples of actions/activities, emotions and social relations.

Dimension	Explanation of different levels
Valance	<i>level -2</i> : obviously negative; <i>level -1</i> : general rather than obviously negative; <i>level 0</i> : neutral; <i>level +1</i> : no negative emotion when talking with others; <i>level +2</i> : obviously positive
Arousal	<i>level -2</i> : lying, sitting, or standing still with no movement of the hands; <i>level -1</i> : sitting or standing with slight movement of the hands; <i>level 0</i> : walking; <i>level +1</i> : simple sport (e.g., jogging and morning exercise); <i>level +2</i> : sports or exercise with stretched limbs
Dominance	<i>level -2</i> : lying or sitting; <i>level -1</i> : standing; <i>level 0</i> : walking; <i>level +1</i> : doing sport or exercise; <i>level +2</i> : energetic exercise (e.g., , running)

Table 4. Explanations of emotion levels in terms of specific scenarios in the public space.

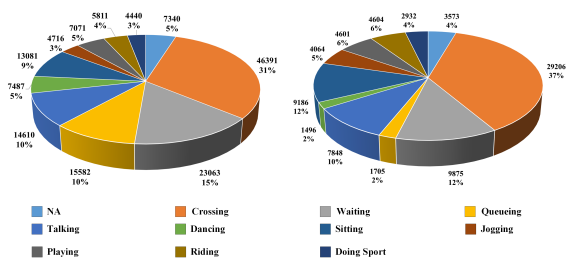


Figure 2. Statistics of actions (left) and activities (right).

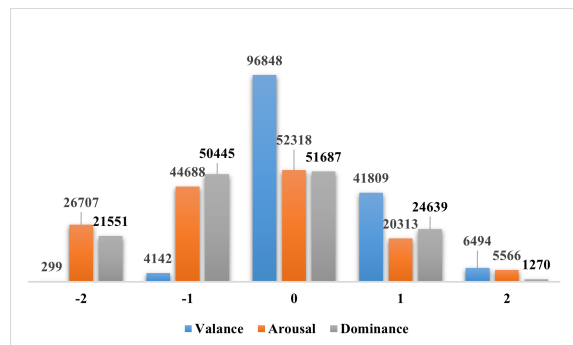


Figure 3. Individual statistics of emotion recognition.

5. Dataset Design

Data are foundational for the design and training of deep learning networks while this is far from enough. How

	-2	-1	0	+1	+2	NA	Cro.	Wai.	Que.	Tal.	Dan.	Sit.	Jog.	Pla.	Rid.	Do.
V	0.91	0.77	0.92	1.00	0.80	0.70	0.91	0.88	0.83	0.97	1.00	0.90	0.94	0.81	0.94	1.00
A	0.93	0.93	0.77	0.92	0.81	Fri.	Fam.	Cou.	Pro.	Com.	No.	-	-	-	-	-
D	0.90	0.89	1.00	0.87	0.90	0.96	0.88	0.94	1.00	1.00	0.97	-	-	-	-	-

Table 5. Consistency analysis of actions/activities, emotions (valance, arousal and dominance) and social relations.

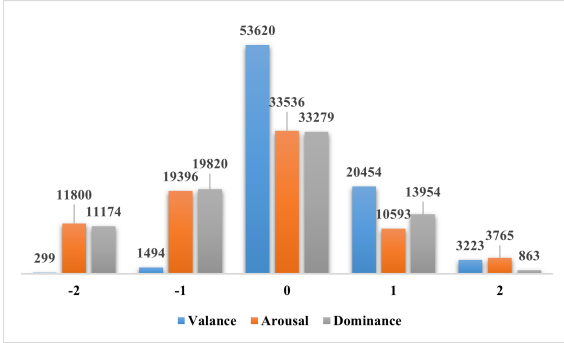


Figure 4. Group statistics of emotion recognition.

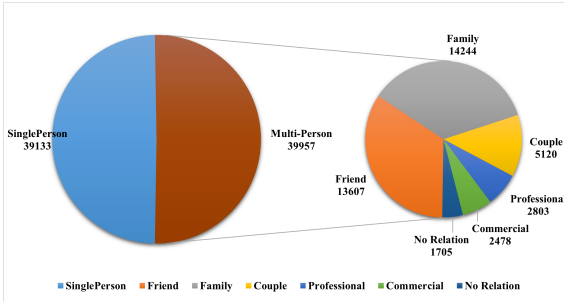


Figure 5. Statistics of social groups (including single-person groups and multi-person groups).

to sample and split is the key factor for the effectiveness and the robustness of the networks.

According to different tasks (i.e., action/activity recognition, emotion recognition and SRR), we adopt different sampling strategy and split methods to obtain the training set and testing set.

For the video-based action/activity recognition, we randomly split the 71 clips into the training set (56 clips) and testing set (15 clips), which is a simple but effective method. For the emotion recognition, we split video segments into training set and testing set, which are cut from the 71 clips and 10 supplementary clips only for algorithm research. For SRR, we split different training sets and testing sets for the pair-based (i.e., social relations between two persons) and group-based (i.e., social relations of the groups with more than two persons) SRR.

	Training	Testing
SRR-Pair	14,258	2,616
SRR-Group	13,462	2,425
Emotion-Individual	4,173	1,070
Emotion-Group	1,632	384
Individual Action	121,353	28,239
Group Activity	63,334	15,756

Table 6. Split of dataset.

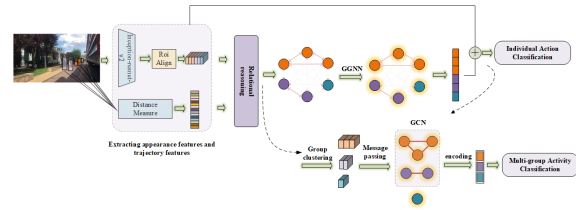


Figure 6. An end-to-end network for multi-group activity recognition, including group division, individual action recognition and group activity recognition.

6. Baseline Methods

To promote the development of traditional tasks, we also present the baselines of action recognition, individual emotion recognition and SRR based on person-pair besides multi-group activity recognition, group emotion recognition and group SRR. The split of training and testing sets is listed in Table 6. The code and dataset with the annotations for training and testing will be made freely available to academic and non-profit organizations for non-commercial, scientific use under the premise of privacy protection mentioned in Section 2.

6.1. Multi-Group Activity Recognition with Individual Action Recognition

There are usually multiple groups in one scene, so we proposed an end-to-end network for multi-group activity recognition with individual action recognition.

As shown in Figure 6, the network consists of three parts, including the individual feature extraction (IFE) module, distance feature extraction (DFE) module, group clustering module and graph reasoning module. Individuals' features are extracted by the Inception-ResNet-v2 [37] and RoI-Align module [12] and the DFE module calculates the

Euclidean distance of each two individuals as the distance features by feeding the coordinates of bounding boxes. For the group clustering module, all individuals are divided into different groups through correlations (cosine similarity or Pearson coefficient) of each pair based on the extracted individuals' features and the calculated distance features. At the stage of graph reasoning, two graph structures are constructed for individual action classification and multi-group activity classification while the reasoned individual action features are used to infer group activity. Specifically, we treat the individuals' features as the node and connect them with the correlations of each pair to form graph G_1 . Then gated graph neural network (GGNN) [25], emphasizing message propagation between nodes, is introduced to reason this graph for reasoned individuals' features. And finally, the reasoned individuals' features, the output of GGNN, concatenate the individuals' features extracted by the IFE module for individual action classification. For the multi-group activity classification, the reasoned individuals' features are viewed as the node and their connections are up to the group division. That is, nodes only in the same group are connected to each other so that a group corresponds to a sub-graph for the entire graph structure G_2 . After graph construction, graph convolutional network (GCN) [14], concentrating on topological structure, is introduced to reason this graph for the final multi-group activity classification.

The performance of multi-group activity recognition is related to group clustering and individual action recognition, so we design a multiple loss to optimize our model. It is composed of three-loss functions:

$$L = \lambda_1 \sum_s L_{sgp} (O_s^{SG}, \hat{O}_s^{SG}) + \lambda_2 \sum_n L_{ind} (O_n^I, \hat{O}_n^I) + \lambda_3 L_c (O^\alpha, \hat{O}^\alpha)$$

where λ_1 , λ_2 , and λ_3 are the balance coefficients, L_{sgp} and L_{ind} , denote the cross-entropy loss functions of multi-group activity recognition and individual action recognition, L_c denotes the binary cross entropy of group clustering, O and \hat{O} denote ground truth and prediction, s and n are the numbers of predicted groups and individuals.

The results of multi-group activity recognition are listed in Table 7. And all metrics of group clustering, individual action recognition, and group activity recognition are denoted as the accuracy, but their computations are different.

1) *Group clustering*. Suppose that there are n people in an image, we use an $n \times n$ matrix M to express group distribution, where $M(i, j) = 1$ denotes the i -th and j -th individual are in the same group and $M(i, i) = 0$ by default. Accordingly, M_g and M_p represent the group distribution of ground truth and prediction, and the accuracy of group clustering can be calculated as follow,

$$Acc_{clu} = \frac{\sum_{i=1}^n \sum_{j=1}^n \delta\{M_g(i,j)=M_p(i,j)\}}{n \cdot n}$$

Group Clustering	Individual Action	Group Activity
64.17	51.07	55.04

Table 7. Results (accuracy) for multi-group activity recognition.

	Valance	Arousal	Dominance
Individual Emotion	48.75	60.22	70.49
Group Emotion	49.87	62.47	64.27

Table 8. Experimental results (accuracy) for individual/group emotion recognition.

$$\delta(x, y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{otherwise} \end{cases}$$

2) *Individual action recognition*. Its accuracy equals the ratio of true positive samples in all the samples.

3) *Group activity recognition*. We view the label of group activity as the one of individual, that is, every individual is tagged an activity label. With individual activity labels, its calculation is same as the one of individual action recognition.

6.2. Individual/Group Emotion Recognition

Data augmentation. We collected some extra videos from the internet for algorithm study. Notably, the collected videos are still from public spaces while their scenarios (e.g., valance level = -2) do not often appear in the public space. Therefore, these videos enrich the minor samples.

Emotion recognition and results. We introduced a pre-trained 3D-ResNet in [11] to exploit the dynamic information of individual patch and group patch. As listed in Table 8, we can observe that the accuracy achieves 48.75%/49.87% for valance, 60.22%/62.47% for arousal, and 70.49%/64.27% for dominance in the individual/group emotion recognition tasks. The slightly low accuracy of valance demonstrates that it is a challenging task to recognize valance levels in the real public spaces (e.g., fuzzy appearance).

6.3. SRR based on Pair and Group

In the two tasks based on pair and group, what we need to focus on is different. For the former, the features from pair and the scene information are important while for the latter, we should pay more attention to the interactions among individuals in the group. Therefore, we proposed two networks based on pair and group, respectively. It is worth noting that sample types of the professional is not enough rich hence we select some pairs from social relation groups to supply pair-based samples of the professional.

Pair-based SRR. Traditional SRR methods [9, 24, 31, 39] have proved that features from person pair and scene

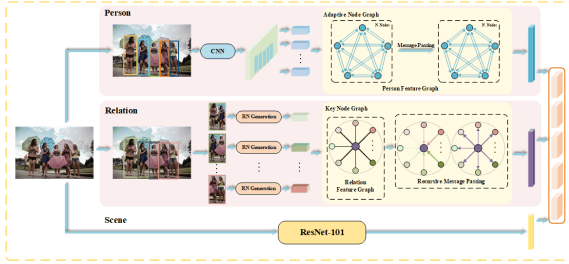


Figure 7. Overall framework of multi-channel SRR network based on interactions among individuals, structure of all relations and scene for group-based SRR.

are effective for SRR. Accordingly, we fuse the two features for SRR. Specifically, we first concatenate the person-pair features (i.e., features of two single body, their union part, and their related position) and then fuse them to obtain a social relation vector representation. Meanwhile, the scene feature is extracted by a pretrained scene model on Places-365-Standard [47] to form a scene vector representation. And finally, the two vector representations are concatenated for recognizing social relations.

Group-based SRR. Compared with the traditional pair-based SRR, group-based SRR has some problems, such as unfixed number of people in the group and indistinguishable individual features, which increases the difficulty for recognition. Therefore, we proposed a multi-channel SRR network based on interactions among individuals, structure of all relations and scene.

As shown in Figure 7, our model contains three channels, which extract the interactions among individuals, the structure features of all relations and the scene information. For the first channel, a pretrained ResNet-50 on ImageNet [5] is utilized to extract the feature of each individual in the group and then these features are viewed as node to construct a fully-connected graph, namely person feature graph. This graph contains the interactions among individuals, which can be reasoned by GCN. Thanks to the message propagation, each node contains the information of the other nodes besides itself hence we take one node randomly as the output of this channel. For the second channel, we extract the features of pairs’ union part as relation nodes to construct a relation feature graph for the structure features of all relations. Unlike the person feature graph, a root node is added into this graph and the relation nodes connect to the root node. By this means, the root node will contain the structure features of all relations after the reasoning by GGNN. For the third channel, a scene recognition network, Resnet-101 pretrained on Places365-Standard, extracts the features of full image as the scene information. Finally, the interactions among individuals, the structure features of all relations and the scene information extracted by the three channels are fused to infer social relation together.

Fri.	Fam.	Cou.	Pro.	Com.	No.	mAP
55.5	47.3	39.8	47.4	86.7	78.6	64.1
48.5	54.2	-	95.1	23.3	96.9	71.0

Table 9. Results for SRR based on pair (row 1) and group (row 2).

Same as [24], we present the per-class recall and mean average precision (mAP) as evaluation metrics of our SRR model in Table 9.

7. Conclusion and Future Work

In this paper, we provide a new multi-group video dataset (PLPS) from the real public space for public life research and affective behavior analysis with annotations of three tasks. This is aimed to promote the applications of computer vision on the comprehensive affective behavior analysis for HCI systems in the wild. The proposal of this dataset will extend the studies of three tasks in the real public space based on multiple groups. Especially, the baseline models of multi-group activity recognition and group-based SRR were proposed to show their feasibility. In the future, we plan to design a uniform framework for these three tasks. And then based on low-level basic information (e.g., age, gender, and group), higher level attributions from the uniform framework and learned correlations among them, we also plan to deepen the public life understanding for affective behavior analysis and further boost the development of HCI systems.

Acknowledgement

This research was supported by the National Nature Science Foundation of China under Grant 61871278 and the Sichuan Science and Technology Program under Grant 2018HH0143.

References

- [1] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4), 2008.
- [2] Wongun Choi and Silvio Savarese. A unified framework for multi-target tracking and collective activity recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 215–230, 2012.
- [3] Wongun Choi, Khuram Shahid, and Silvio Savarese. What are they doing?: Collective activity classification using spatio-temporal relationship among people. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 1282–1289, 2009.
- [4] Wongun Choi, Khuram Shahid, and Silvio Savarese. Learning context for collective activity recognition. In *Proceed-*

- ings of the *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3273–3280, 2011.
- [5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.
- [6] Abhinav Dhall, Roland Goecke, Simon Lucey, and Tom Gedeon. Collecting large, richly annotated facial-expression databases from movies. *IEEE MultiMedia*, 19(3):34–41, 2012.
- [7] Ruogu Fang, Andrew C. Gallagher, Tsuhan Chen, and Alexander Loui. Kinship classification by modeling facial feature heredity. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 2983–2987, 2013.
- [8] Jan Gehl and Birgitte Svarre. *How to study public life*. Island press, 2013.
- [9] Arushi Goel, Keng Teck Ma, and Cheston Tan. An end-to-end network for generating social relationship graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11178–11187, 2019.
- [10] Yuanhao Guo, Hamdi Dibeklioglu, and Laurens Van Der Maaten. Graph-based kinship recognition. In *Proceedings of the International Conference on Pattern Recognition*, pages 4287–4292, 2014.
- [11] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555, 2018.
- [12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017.
- [13] Mostafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, and Arash Vahdata. Social roles in hierarchical models for human activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1354–1361, 2012.
- [14] Thomas Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.
- [15] Dimitrios Kollias, Irene Kotsia, Elnar Hajiyev, and Stefanos Zafeiriou. Analysing affective behavior in the second abaw2 competition. *arXiv preprint arXiv:2106.15318*, 2021.
- [16] Dimitrios Kollias, Mihalis A Nicolaou, Irene Kotsia, Guoying Zhao, and Stefanos Zafeiriou. Recognition of affect in the wild using deep neural networks. In *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017 *IEEE Conference on*, pages 1972–1979. IEEE, 2017.
- [17] D Kollias, A Schulc, E Hajiyev, and S Zafeiriou. Analysing affective behavior in the first abaw 2020 competition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)(FG)*, pages 794–800.
- [18] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019.
- [19] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multi-task learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021.
- [20] Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arface. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 78.1–78.15, September 2019.
- [21] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021.
- [22] Ronak Kosti, Jose M. Alvarez, Adria Recasens, and Agata Lapedriza. Context based emotion recognition using emotic dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(11):2755–2766, 2020.
- [23] Jiyoung Lee, Seungryong Kim, Sunok Kim, Jungin Park, and Kwanghoon Sohn. Context-aware emotion recognition networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 10142–10151, 2019.
- [24] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S. Kankanhalli. Dual-glance model for deciphering social relationships. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2669–2678, 2017.
- [25] Yujia Li, Richard Zemel, Marc Brockschmidt, and Daniel Tarlow. Gated graph sequence neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.
- [26] Xinchun Liu, Wu Liu, Meng Zhang, Jingwen Chen, Lianli Gao, Chenggang Yan, and Tao Mei. Social relation recognition from videos via multi-scale spatial-temporal reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3561–3569, 2019.
- [27] Jiwen Lu, Xiuzhuang Zhou, Yap-Pen Tan, Yuanyuan Shang, and Jie Zhou. Neighborhood repulsed metric learning for kinship verification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(2):331–345, 2014.
- [28] Patrik Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *Proceedings of Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 91–101, 2010.
- [29] Jinna Lv, Wu Liu, Lili Zhou, Bin Wu, and Huadong Ma. Multi-stream fusion model for social relation recognition from videos. In *Proceedings of the International Conference on Multimedia Modeling (MMM)*, pages 355–368, 2018.
- [30] A Mehrabian. Constants across cultures in the face and emotion. *Genetic, Social, and General Psychology Monographs*, 121(3):339–361, 1995.
- [31] Meng Zhang; Xinchun Liu; Wu Liu; Anfu Zhou; Huadong Ma; Tao Mei. Multi-granularity reasoning for social relation recognition from images. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 1618–1623, 2019.

- [32] Trisha Mittal, Pooja Guhan, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emoticon: Context-aware multimodal emotion recognition using frege’s principle. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14222–14231, 2020.
- [33] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2019.
- [34] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Proceedings of the International Conference on Multimedia and Expo (ICME)*, pages 317–321, 2005.
- [35] Harry Reis, Willard Collins, and Ellen Berscheid. The relationship context of human behavior and development. *Psychological Bulletin*, 126:844–72, 12 2000.
- [36] Qianru Sun, Bernt Schiele, and Mario Fritz. A domain based approach to social relation recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 435–444, 2017.
- [37] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, page 4278–4284, 2017.
- [38] Gang Wang, Andrew Gallagher, Jiebo Luo, and David Forsyth. Seeing people in social context: Recognizing people and social relationships. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 169–182, 2010.
- [39] Zhouxia Wang, Tianshui Chen, Jimmy Ren, Weihao Yu, Hui Cheng, and Liang Lin. Deep reasoning with knowledge graph for social relationship understanding. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1021–1028, 2018.
- [40] Siyu Xia, Ming Shao, and Yun Fu. Kinship verification through transfer learning. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2539–2544, 01 2011.
- [41] Haibin Yan and Junlin Hu. Video-based kinship verification using distance metric learning. *Pattern Recognition*, 75:15–24, 2018.
- [42] Rui Yan, Lingxi Xie, Jinhui Tang, Xiangbo Shu, and Qi Tian. Social adaptive module for weakly-supervised group activity recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 208–224, 2020.
- [43] Senthil Yogamani, Ciaran Hughes, Jonathan Horgan, Ganesh Sistu, Sumanth Chennupati, Michal Uricar, Stefan Milz, Martin Simon, Karl Amende, Christian Witt, Hazem Rashed, Sanjaya Nayak, Saquib Mansoor, Padraig Varley, Xavier Perrotton, Derek Odea, and Patrick Pérez. Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9307–9317, 2019.
- [44] Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A. Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-wild: Valence and arousal ‘in-the-wild’ challenge. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1980–1987, 2017.
- [45] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. From facial expression recognition to interpersonal relation prediction. *International Journal of Computer Vision*, 126(5):550–569, 2016.
- [46] Guoying Zhao, Xiaohua Huang, Matti Taini, Stan Z. Li, and Matti Pietikäinen. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9):607–619, 2011.
- [47] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2018.