Relationship existence recognition-based social group detection in urban public spaces

Lindong Li [a], Linbo Qing [a,*], Li Guo [b], Yonghong Peng [b]

[a] College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China
[b] Department of Computing and Mathematics, Manchester Metropolitan University, Manchester M1 5GD, UK

A B S T R A C T

In urban public spaces, a social group consists of two or more individuals who share some social relationships and interact based on mutual expectations. However, most existing studies found people's F-formations on a top view, which is hard to observe their social contexts and the top-view videos are not easily accessible in real urban life. Recently, some researchers turned to urban scenes and analysed front-view human behaviours for social group detection. But these methods still cannot grasp the nature of social groups, i.e., the relationships among individuals. It is the key to finding social groups to judge whether any two individuals belong to the same cluster. Therefore, this paper proposes a new paradigm: relationship existence recognition-based social group detection. Additionally, on top of the paradigm, we designed a new social group detection algorithm incorporated with the visual cue-based and non-visual cue-based components. Specifically, the former exploits the spatial interactions and the temporal information to recognise the existence of social relationships through supervised deep learning. The latter estimates the similarities of trajectory pairs using the unsupervised spatial–temporal position information. Social group detection achieves superior accuracy with the two components' complementary results. On Social-CAD (Social Collective Activity Dataset) and PLPS (Public Life in Public Space) datasets, extensive experiments demonstrate that our algorithm outperforms the state-of-the-art (SOTA) methods.

1. Introduction

In the urban research community, it is an important investigation method to observe citizens and their activities from the perspective of human beings [1,2]. This observation can guide the planning, construction and management of a people-oriented city. It is essential to develop a computer vision technique for finding social groups in urban public spaces to achieve this goal. However, few methods are robust to this scene setting. The following will give the related introduction to social group detection and analyse the challenges based on existing studies.

Social group detection aims to find clusters of two or more individuals. In such clusters, individuals share some social relationships, e.g., friends, family, and professionals; hence it can be called a social group. Noting that detection task is only to group individuals and recognising their social relationships is another field. Existing methods of social group detection can be categorised into two paradigms, namely positional structure-based methods [3,4] and human behaviour-based methods [5,6]. As shown in Fig. 1, the former detects the F-formation [7,8] structures to find social groups according to the individual positions and head orientations. Generally, an F-formation structure comprises three spaces, i.e., o-space, p-space, and r-space. The essential o-space is shared by all group members yet cannot access by anyone, so most researchers focused on detecting this space. The latter emphasises the pre-defined individual behaviours, such as waiting, walking, and talking. Usually, it incorporates the study of collective activity recognition, where the group detection module utilises the semantic context of individual behaviours to find clusters.

Many efforts have been made in the past decades based on the above two paradigms. However, there are still some limitations. As for the position structure-based methods, the individual positions and head orientations must be utilised on a top view. It restricts the applications in urban public spaces, where human beings are observed from a front view [9,1,2]. Besides, these methods mainly detect the standing conversational groups, lacking robustness to other types of social groups. To be polite, the standing conversational group members meet face to face and form a fixed F-formation. However, in some other groups (e.g., the dawdling group), the members' orientations may go as they please. Under this condition, these methods are hard to perform well even

Fig. 1. Positional structure-based (left) and human behaviour-based (right) social group detection.

though the top-view F-formations could be inferred from the front-view videos. For the human behaviour-based practices, the predefined behaviours cannot cover all types, especially for some middle actions (e.g., from standing to sitting), which are difficult to describe and define. In some social groups, behaviours differ individuals in the same cluster. It can be easy to imagine a scenario: one is standing and talking, but the others are sitting and listening. Though the human behaviour-based methods made some achievements from the front view, it is hard to detect social groups accurately, only according to the individual behaviours or their semantic context.

In a word, the two paradigms cannot hit the nature of social groups, which causes the above limitations in urban public spaces. In social groups, social relationships exist, which are understood as the set of connections among group members. These connections determine who belongs to the same social group and their types reflect the attributes of the social groups. For example, we establish affective and labour relationships to form friend groups and professional groups, respectively. In other words, the social relationships among group members are the nature of social groups. For the social group division or detection, the core problem is to judge whether individual pairs share connections, i.e., social relationships. It differs from the two existing paradigms and can be viewed as a binary classification problem, namely relationship existence recognition.

Therefore, based on the new paradigm, this paper proposed a corresponding algorithm incorporating the visual and non-visual cues. Firstly, we exploit the visual interactions to recognise the relationship's existence between individuals using supervised deep learning. Notably, the algorithm focuses on spatial and temporal cues to avoid confusion between the focused and unfocused encounters [8,10]. We also designed a feature extraction mechanism for spatial cues to obtain the multilevel semantic information from a pre-trained model. It can enhance the feature representations of fuzzy appearances in urban public spaces. Secondly, the similarity of the trajectory pair is an important non-visual cue, so we also measure it to recognise the relationship's existence. Finally, the recognition results from the visual and non-visual cues are fused complementarily for better performance of social group detection.

The main contributions of this paper are summarised as follows:

1. A new paradigm based on social relationships is proposed for social group detection. It first shifts the complex problem into the simple yet effective binary classification, i.e., relationship existence recognition. It hits the nature of social group detection, which can be described as the relationship between individuals determining whether they belong to the same social group. Meanwhile, it provides the primary knowledge of interpersonal relationships and bridges social group detection and relationship understanding.
2. A new algorithm based on the above paradigm is proposed. It complementarily fuses the spatial–temporal information of the interactions and the positions. They are extracted from the visual patches of person pairs and the non-visual similarities of trajectory pairs using supervised deep learning and unsupervised index measurements.
3. A novel multilevel feature extraction (MFE) mechanism is designed to extract comprehensive semantic information from a pre-trained model. It can enhance the feature representations and weaken the side effect of fuzzy appearances in urban public spaces.
4. Comprehensive experiments have been conducted on PLPS (Public Life in Public Space) and Social-CAD Social Collective Activity Dataset datasets. The algorithm outperforms the state-of-the-art (SOTA) methods, demonstrating our algorithm's superiority in public spaces. Meanwhile, an ablation study was also conducted to prove the effectiveness of visual and non-visual cues.

The rest of this paper is organised as follows. Section 2 reviews the literature of related work. Section 3 formulates the problem of social group detection based on the new paradigm. Next, the corresponding algorithm is described in Section 4. On public datasets, experiments are implemented, and results are analysed in Section 5. In the end, Section 6 concludes the paper.

2. Related work

This paper aims to solve the problem of social group detection based on the existence recognition of social relationships. Hence, this section reviews the related works from the following aspects, i.e., social group detection and social relationship understanding. To further state the significance of the MFE mechanism, we will also introduce the related feature extraction based on pre-trained models.

2.1. Social group detection

In the last two decades, researchers detected the social groups based on positional structure and human behaviour, respectively [11].

The positional structure-based social group detection was derived from a sociological notion, F-formation [7,8]. An F-formation consists of the individuals and the spatial pattern, determined by their positions and orientations. Based on this knowledge, researchers in the computer vision field started to detect the F-formation structure for finding the social groups in the crowd. Yu et al. [12] proposed the modularity-cut algorithm to discover groups and their leadership structures. Cristani et al. [13] and Hung et al. [10] first introduced the F-formation and designed algorithms for social group detection. Bazzani et al. [14] estimated attention's visual focus and proposed an inter-relation matrix to suggest possible social interactions. Cristani et al. [13] utilised Hough voting scheme to determine the o-space based on the individual position and head orientation. Hung et al. [10] calculated the affinities of person pairs for the edge-weighted graph and used a graph clustering algorithm for identifying dominant sets as F-formations. It is worth noting that Tran et al. [15] used the same

graph clustering algorithm for group detection and activity recognition. Following this line, Setti et al. [16] summarised the strengths and weaknesses of Hough voting scheme and graph clustering algorithm. Besides, they pointed out that the former could resist the noise using head orientation information while the latter had better performance with only position information available.

Furthermore, Setti et al. extended the Hough voting scheme for a multi-scale F-formation detection [17] and proposed a graph-cuts based framework for clustering individuals [18]. Yasuda et al. [19] proposed observing lower bodies to describe F-formations further. Zhang et al. [20] developed an extensive study of social involvement and proposed to detect associates [21] of F-formations. In addition to the above studies in still images, some researchers introduced extra temporal information to enhance the task of social group detection. Gan et al. [22] modelled the temporal information for an extended F-formation system with the function of social interaction detection. Vascon et al. [23,24] embedded the temporal constraints into a game-theoretic framework to check the head orientation and pose estimation. Inaba et al. [3] considered the individual visual attention field changes and presented the robust detection method for time-varying F-formation. Cabrera-Quiros et al. [25] collected a multi-sensor dataset to analyse social interactions and group dynamics, involving videos. In recent years, various methods have emerged for social human-robot interactions. Pathi et al. [26] proposed a real-time algorithm to estimate the face orientation and detect the F-formations, expecting to boost the human-robot interactions.

Consequently, Pathi et al. [27] presented a model to find the o-space and estimated the optimal placement for a robot. Pathi et al. [9] also addressed the problem of optimal placement estimation under the circumstance of multiple groups from an ego-view. Besides, Barua et al. [4] extended to predict the robot's path angle for joining the social group, depending on the real-time F-formation recognition.

Compared to the above F-formation-based methods with a long history, human behaviour-based methods had increasing attention in multi-group activity recognition. After decades of studying individual action and single group activity, the computer vision community recently has started concentrating on the activity understanding of different groups in the same scenario. In this field, as the sub-task and pre-task, social group detection usually is learned and incorporated with the end-to-end framework with multiple tasks, i.e., group division, individual action classification and sub-group activity recognition. Ehsanpour et al. [5] annotated different social groups and the corresponding social activity labels to extend the original collective activity dataset [28]. In their proposed multi-group activity recognition framework, individual behaviour features are extracted as nodes and graph spectral clustering is used to divide social groups. Qing et al. [6] collected a new dataset to sense the public life in public spaces. They proposed their framework, in which individual behaviour features are extracted, and the individuals with high similarity are grouped into one cluster.

The above descriptions review the literature on social group detection. However, as mentioned in Section 1, the methods are not robust to the urban public spaces andthe two paradigms cannot hit the nature of social groups. They focused on the manifestations, including positions, orientations, and behaviours. Therefore, this paper proposes the relationship existence recognition-based method, exploiting the social cues of relationships among individuals for social group detection.

2.2. Social relationship understanding

In computer vision, relationship understanding can be summarised as three main aspects, involving relative position among objects [29,30], dominant action of the person over object [31–33], and interactive behaviour between persons [34,35]. Compared with these intuitive aspects, a social relationship is defined abstractively based on the theory of sociology and psychology [36,37]. Specifically, it refers to the links among individuals in a social group. Understanding the existence of these links is the key problem in judging whether the two individuals belong to the same social group. According to the research targets, scholars mainly focus on two aspects: social relationship recognition [38–42] and social network generation [43,44]. The former recognises the specific types of social relationships while the latter aims to find the different camps in films and TV shows. For example, Yang et al. [40] derived inspiration from studying human gaze communication [45,46]. They proposed a gaze-aware graph convolutional network to recognise social relations, e.g., friends, family, and colleagues. Lv et al. [44] explored video and subtitle text information to form a relationship network, which is analysed to discover communities and important roles in the story.

In this field, most works contribute to verifying the specific social relations and only limited efforts are made for social network generation. However, these works mainly understand social relationships based on relationship existence. From the perspective of datasets, most of them only consist of one person pair or one social group in an image or video clip, e.g., [47], IRD [48], PIPA [49], and ViSR [50]. In other words, "no relation" hardly exists in these datasets. Even if "no relation" is labelled in PISC [51] dataset and PLPS [6] dataset, their baselines still emphasise the recognition of specific social relations in images. The existence verification of the relationship is weak, and the temporal information is ignored, which is vital for social group detection.

To summarise, the existing studies of social relationship understanding cannot meet the needs of our proposed paradigm. Hence, designing the corresponding algorithms for social relationship understanding and social group detection is pretty meaningful.

2.3. Feature extraction based on pre-trained models

In machine learning, it is a popular practice to use pre-trained models for feature extraction, which falls under the category of transfer learning. Usually, they are trained on an extensive dataset and utilised to solve problems on another dataset based on deep learning. Over the past two decades, data-driven artificial intelligence has boomed and researchers have collected many large-scale datasets, e.g., ImageNet [52], Places365 [53] and Kinetics [54]. Consequently, various models, e.g., ResNet [55], ViT [56] and BERT [57], are pre-trained on these datasets and used for other tasks [39,58–60].

However, most existing methods only utilised the final feature representations (i.e., the highest semantic features) as the input of downstream tasks. They ignored the lower-level information. Meanwhile, fuzzy appearances in urban public spaces are challenging for feature extraction and the final feature representations are insufficient. Hence, this paper proposed a MFE mechanism to obtain comprehensive semantics from pre-trained models.

3. Problem formulation

This paper proposes a new paradigm of social group detection, namely the relationship existence recognition-based method. This paradigm aims to group individuals who share some social relationships in a given urban public space video. The following will formulate the general flow of our proposed paradigm. Let us first define the set of individuals in a video sequence as $I = \{x_1, x_2, \ldots, x_N\}$, where N denotes the total number of individuals in the space. Then, our key problem is to judge whether the social relationship exists for every person pair or not, i.e., relationship

existence recognition. Let $M \in \mathbb{R}^{N\times N}$ denotes the recognition results, where M_{ij} equals to 1 if the i-th individual shares a social relationship with the j-th one otherwise 0. It can be formulated as follows,

$$M_{ij} = M_{ji} = f(x_i, x_j), M = \{M_{ij} | 1 \leqslant i \leqslant N, 1 \leqslant j \leqslant N\} \quad (1)$$

where $f(\cdot)$ denotes the function of relationship existence recognition. Because both M_{ij} and M_{ji} denote the relationship existence between the i-th individual and the j-th one, M is a symmetric matrix, i.e., $M_{ij} = M_{ji}$. In addition, the relationship between one individual and themselves is meaningless, so M_{ii} or M_{jj} equals zero as default. Finally, we can map the result matrix M into social groups based on the social relationship's transitivity rule $g(\cdot)$. That is to say, if A shares a relationship with B and C, then so do B and C.

Let us take Fig. 2 as an example to visualise the flow more clearly. In the left picture, there are $N = 6$ individuals with labelled bounding boxes, who can be first initialised as $I = x_1, x_2, \ldots, x_6$ from left to right. Then, according to the function $f(\cdot)$ of relationship existence recognition, every person pair can be judged as "has relation" or "no relation" and the results can be shown as the adjacent matrix. In this example, there are four person pairs with relationships, i.e., $x_1 - x_2, x_1 - x_3, x_2 - x_3$, and $x_5 - x_6$. Correspondingly, $M_{12} = M_{21} = 1, M_{13} = M_{31} = 1, M_{23} = M_{32} = 1$, and $M_{56} = M_{65} = 1$, and the other elements are zero. It is worth noting that x_4 does not share a relationship with any others, and he is a single-body individual, which is viewed as a special social group in this paper. Finally, draw lines to illustrate the relationships among individuals. In the topological graph, nodes and edges represent the individuals and relationships. It can be observed that there are three social groups, i.e., $\{x_1, x_2, x_3\}, \{x_4\}$, and $\{x_5, x_6\}$. The visualised sample is shown in the right picture.

Throughout the entire flow, there is no doubt that the relationship existence recognition $f(\cdot)$ is the key to the paradigm. Hence, the following sections will elaborate on relationship existence recognition, including the supervised and unsupervised components.

As for the transitivity rule $g(\cdot)$, it is a simple yet important process method, which utilises the logical constraints among social relationships of person pairs at the group level. As shown in Fig. 3, take four people (A, B, C, and D) for example. From the perspective of social relationship detection for every person pair, it is easy to determine the relationship existence of (A,B), (A,C), and (A, D) because of their gaze communications. With this limited information, (A,B,C,D) can also be inferred as a social group even though we do not know the relationship existence of the person pair (B,C), (B,D), and (C,D). In this inference, (B,C), (B,D), and (C,D) are reasoned as person pairs with a social relationship based on the logical constraints. It is worth noting that this group-level logical constraint is proved in the field of social relation recognition [61].

4. Method

As mentioned above, the key problem for the proposed paradigm is to judge whether the social relationship exists for every person pair or not. Therefore, this section will show how our algorithm recognises person pairs' relationship existence. As shown in Fig. 4, the proposed algorithm first infers the existence of a relationship through visual and non-visual cues. Then the fusion module combines the complementary results for the final prediction. The following will first elaborate on the three parts, and then introduce the relation classification and the model optimisation.

4.1. Visual cue-based component

According to traditional feature extraction of social relationship recognition and the real situations of urban public spaces, we extract the sequential temporal dynamics and spatial social interactions to predict social relationships.

4.1.1. Temporal information extraction

To avoid the confusion between the focused and unfocused encounters [8,10], we must consider the temporal information. Hence we select three frames of the whole sequence to represent the temporal information. The classic yet effective combination of "CNN + RNN" is designed to explore it. We first utilise a pre-trained ResNet-101 [55] on ImageNet [52] to extract the features of the union patches cropped from the selected frames, respectively. Given the bounding boxes of two individuals, the union patch refers to their union part, as shown in Fig. 6. Then, these features are fed into the long-short term memory (LSTM) [62] network. Finally, its output is concatenated as the two individuals' temporal information $F_{sd} \in \mathbb{R}^{1536}$. Suppose that the input frames are $X = \{X_1, X_2, \ldots, X_n\}$, then the temporal information extraction can be expressed as follows,

$$T_i = f_{res}(X_i), i = 1, 2, \ldots, n \quad (2)$$

$$F_{sd} = f_{lstm}(T_1, T_2, \ldots, T_n) \quad (3)$$

where $T_i \in \mathbb{R}^{512}$ denotes the features extracted from the i-th union patch, $f_{res}(\cdot)$ denotes the ResNet-101 model and $f_{lstm}(\cdot)$ denotes the LSTM model.

4.1.2. Social interaction extraction

To alleviate the effect of fuzzy appearance in urban public spaces, we further introduce the MFE mechanism for enhancing semantic feature representations of social interactions in this section. As shown in Algorithm 1, given a pre-trained model *Net* with multiple stacked layers or blocks, its multilevel features T_i are mapped into S_i by a set of extraction functions F. The final concatenation result S_{out} contains rich semantic information from low to high levels.

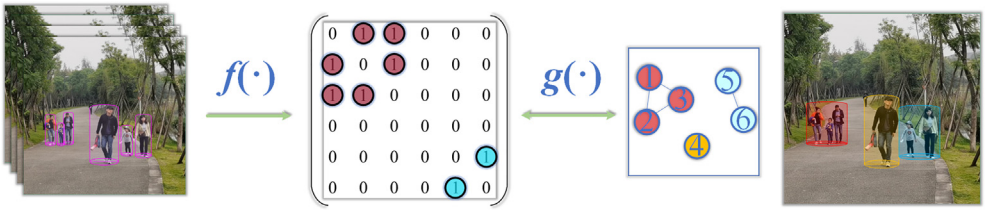

Fig. 2. The visualised flow of the relationship existence recognition-based social group detection.

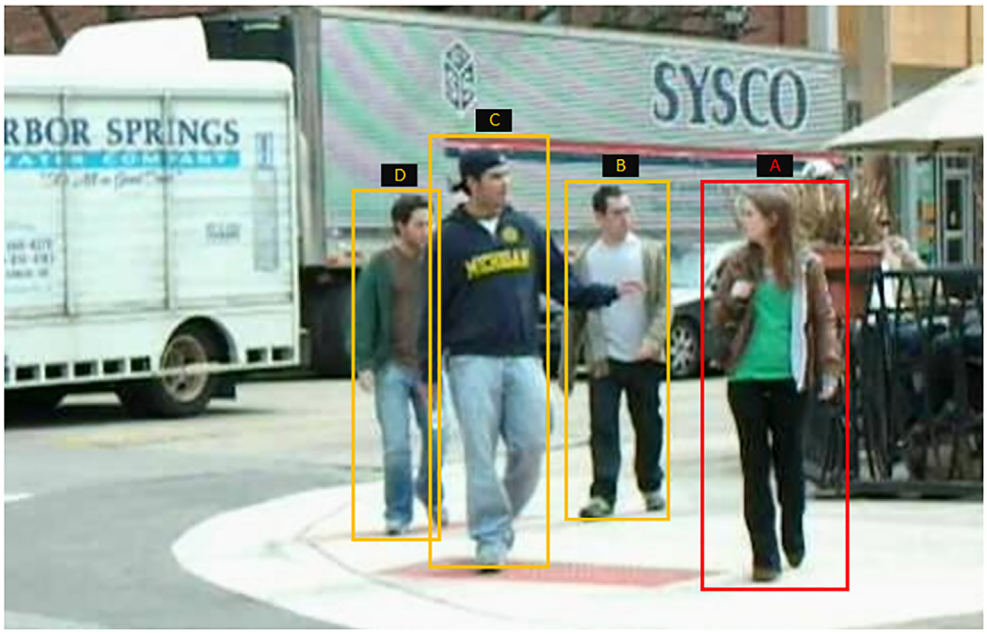

Fig. 3. An example of the group-level logical constraints.

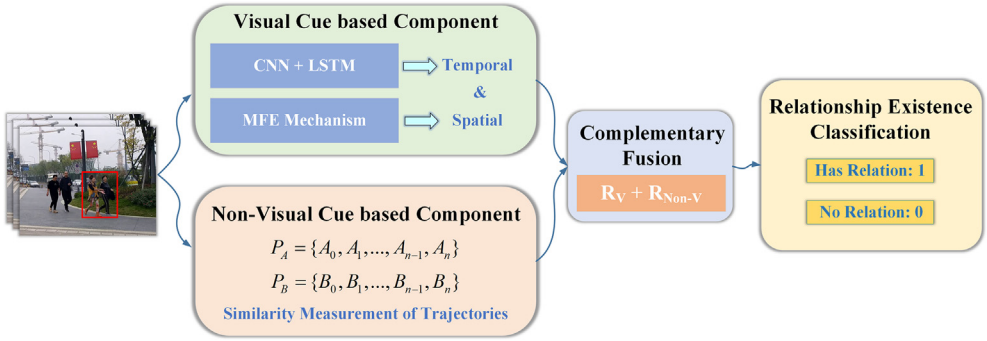

Fig. 4. The overall framework of relationship existence recognition. The visual cue-based component exploits temporal-spatial information using the "CNN + LSTM" combination and the MFE mechanism. In the non-visual cue based component, the similarities of every trajectory pair are measured as extra information. Finally, this information is complementarily fused to classify the existence of social relationships.

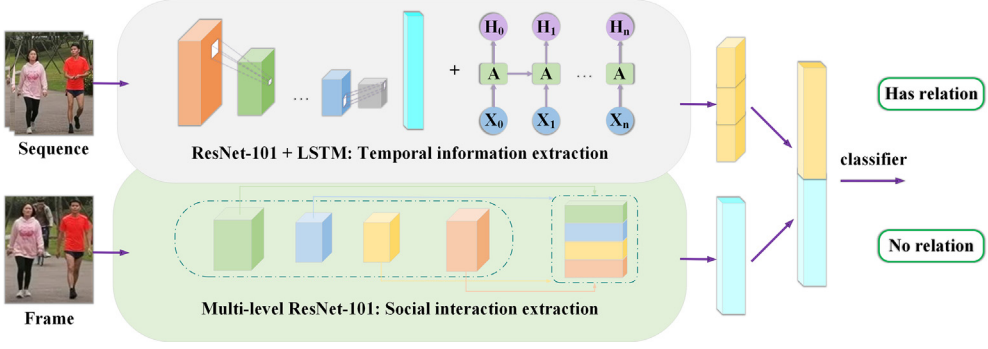

Fig. 5. Visual cue-based component.

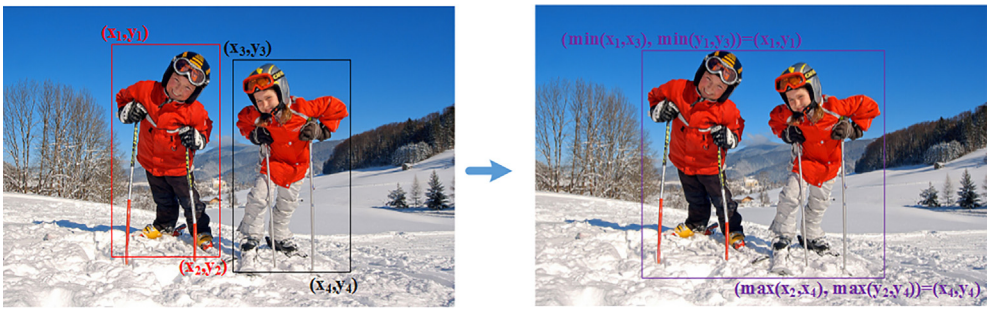

Fig. 6. Union patch cropped from an image/frame. This figure is from [39].

Algorithm 1 MFE: Multilevel feature extraction mechanism

Require: Initialise input: S_{in}
Require: Initialise pre-trained model: $Net = \{b_1, b_2, \ldots, b_n\}$
Require: Initialise extraction function: $F = \{f_1, f_2, \ldots, f_n\}$
Ensure: Feature Representations: S_{out}
1: $T_1 = f_1(S_{in})$;
2: **for** $i = 2 : n$ **do**
3: $T_{i+1} = b_i(T_i)$;
4: **end for**
5: **for** $i = 1 : n$ **do**
6: $S_i = f_i(T_i)$;
7: **end for**
8: $S_{out} = Concat[S_1, S_2, \ldots, S_n]$;

Based on the above mechanism, we utilise a pre-trained ResNet-101 [55] on ImageNet [52] as the backbone to design a multilevel feature extraction module (i.e., multilevel ResNet-101) for extracting social interactions from the union patch of the two individuals. The union patch, cropped from the entire frame and resized to $\mathbb{R}^{3\times224\times224}$, covers amounts of interactions, which are the most important representation of social relations.

The designed multilevel ResNet-101 is shown in Fig. 7. The ResNet-101 model consists of four blocks, and its extracted features range from low to high level with the stack of blocks. In urban public space videos, the features of personal appearances are fuzzy; hence the multilevel ResNet-101 is designed to enhance the representations of body features. Specifically, the extracted features of four blocks are fetched out and processed to uniform low-level or high-level features, respectively. Finally, these features are concatenated as representations of social interactions. Take Block-1, for example, whose output features are $\mathbb{R}^{256\times56\times56}$. First, adaptive average pooling is utilised to compress the features as $\mathbb{R}^{256\times2\times4}$ The feature maps are flattened into $\mathbb{R}^{2048}$, and finally, the dimension is uniformly reduced to $\mathbb{R}^{512}$ by the fully-connected layer. So, the final output is $\mathbb{R}^{2048}$ from the concatenation of four-level features. When the multilevel ResNet-101 is embedded into the entire network, the final output should be subsampled into $\mathbb{R}^{512}$ before the next operation, which denotes F_{si}.

After the above two features are extracted, they are concatenated to recognise the relationship's existence.

$$F = \{F_{si}, F_{sd}\} \in \mathbb{R}^{512+1536} \tag{4}$$

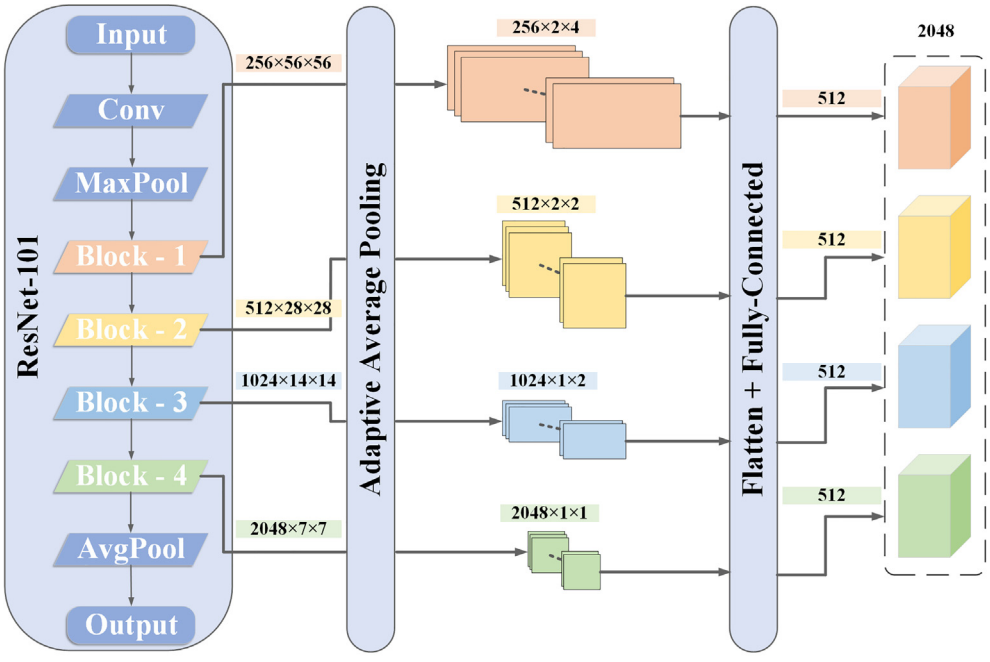

Fig. 7. Multilevel ResNet-101: ResNet-101 with the MFE mechanism.

Fig. 8. Visualised trajectories in urban public spaces.

This component is the essential part of the proposed model, which embodies the main idea, i.e., relationship existence recognition-based social group detection. It is also worth noting that the shared features of ResNet-101 can be reused, promoting the efficiency of forward reasoning. In this module, introducing spatial–temporal information is more important than networks, and other advanced ones can replace them. Hence two-channel architecture without shared features is more general for spatial–temporal information extraction.

4.2. Non-visual cue-based component

Besides the above visual cues, some non-visual cues also indicate the social relationship between individuals. One of the most important is the trajectories, which reflect individuals' various spatial–temporal position information. As shown in Fig. 8, the individual trajectories are drawn in lines with different colours and the social groups are labelled by the red bounding boxes. It can be seen that the trajectories from the same social group are remarkably similar to each other. On the contrary, there are huge differences between the trajectories of different social groups. Therefore, it is essential for the relationship existence recognition to utilise a rational index and to measure the similarities of trajectory pairs.

This paper introduces the Fréchet distance as the similarity degree of trajectory pair. It is a spatial path-based similarity description index proposed by Maurice René Fréchet in 1906, which concentrates on the spatial distance between paths and has high calculation efficiency. Due to the discreteness of video frames, this paper utilises the discrete Fréchet distance to measure the similarity of trajectory pair, which is the approximate calculation of the Fréchet distance.

Before calculating the discrete Fréchet distance, we obtain the trajectories of person pairs by their bounding boxes, which can be expressed as follows,

$$bbox = \{[(p_1, q_1), (m_1, n_1)], [(p_2, q_2), (m_2, n_2)], \ldots, [(p_k, q_k), (m_k, n_k)]\} \tag{5}$$

$$pos_i = \left(\frac{p_i + m_i}{2}, \frac{q_i + n_i}{2}\right), i = 1, 2, \ldots, k \tag{6}$$

$$P = \{pos_1, pos_2, \ldots, pos_k\} \tag{7}$$

where *bbox* denotes the set of individual bounding boxes in k frames, in which $[(p_i, q_i), (m_i, n_i)]$ denotes the coordinates of the top left and bottom right corners of the bounding box in the i-th frame. The final P represents the two-dimension position set, where pos_i denotes the centre coordinates of the bounding box in the i-th frame.

After obtaining the trajectory representations of the person pair, the discrete Fréchet distance can be measured. As shown in Fig. 9, the trajectory pair can be first expressed as follows,

$$P_A = \{A_0, A_1, A_2, \ldots, A_{n-1}, A_n\} \tag{8}$$

$$P_B = \{B_0, B_1, B_2, \ldots, B_{n-1}, B_n\} \tag{9}$$

where P_A and P_B are the instances of trajectory P in Eq. 7, and A_i and B_i are the instances of position pos_i in Eq. 6 simultaneously.

Then, define the distance between P_A and P_B as the maximum of the pairs of positions,

$$\|L\| = \max_{i=0,1,\ldots,n} d(A_i, B_i) \tag{10}$$

where $d(\cdot)$ denotes the Euclidean distance.

Finally, the discrete Fréchet distance can be defined as the minimum $\|L\|$ with the trajectory change,

$$DF(P_A, P_B) = \min \|L\| \tag{11}$$

Denoting that the discrete Fréchet distance is the $\|L\|$ when the trajectory pair is fixed.

After the discrete Fréchet distance of all person pairs is calculated, their corresponding relationship existence can be recognised by setting the threshold τ.

4.3. Complementary fusion

Through the visual and non-visual cues, we can obtain the two adjacent matrixes, i.e., M_{vc} and M_{nvc}, which stand for the results of relationship existence recognition, mapping the social groups as described in Section 3. To fuse them, we calculate the corresponding elements by AND operation,

$$M = M_{vc} \otimes M_{nvc} \tag{12}$$

where $\otimes$ denotes the element-wise product operation and M denotes the final result of social group detection, consisting of 0 and 1. The former means that the person pair does not belong to the same group; otherwise, the latter suggests that the two individuals are in the same group.

4.4. Relation Classification andModel optimisation

The framework comprises two components, i.e., the visual cue-based component and the non-visual cue-based component, corresponding to the supervised deep learning module and the unsupervised unlearned module. Next, we will present their optimisation processes, separately.

As a binary classification model, the output of the visual cue-based component can be expressed as follows,

$$y = \{y_1, y_2\} = FC(M) \tag{13}$$

where *FC* denotes the fully-connected layer.

Then we calculate the probability of each social relationship existence by the SoftMax function as follows,

$$p_i = \frac{e^{y_i}}{e^{y_1} + e^{y_2}}, i = 1, 2 \tag{14}$$

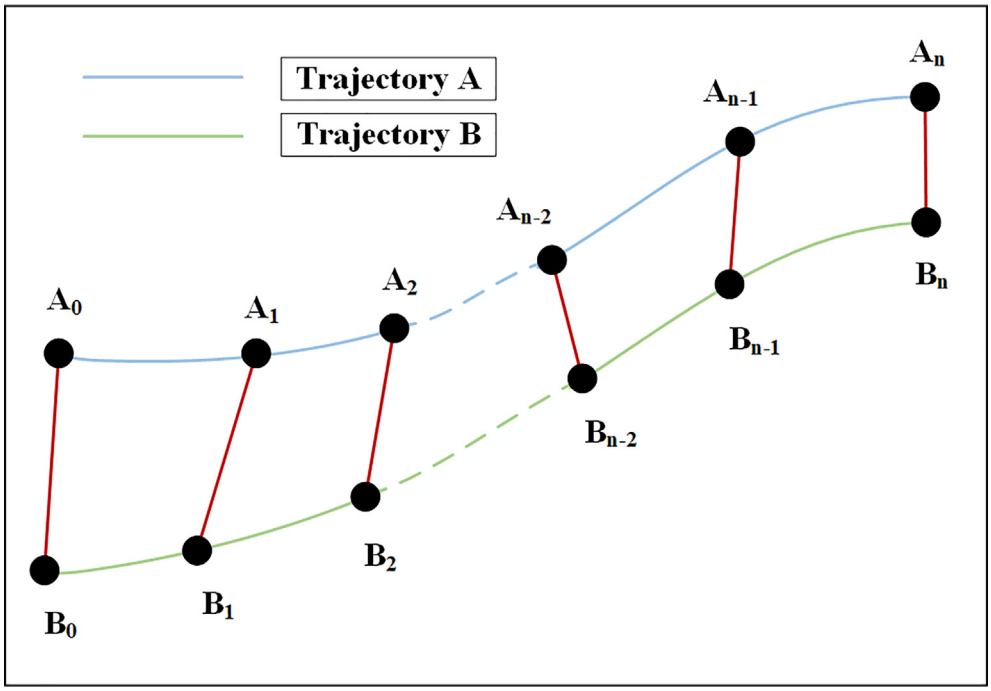

Fig. 9. Trajectory pair for the calculation of the discrete Fréchet distance.

where p_1 and p_2 denote the probabilities of "no relation" and "has relation". The larger probability represents the final classification result.

Finally, according to the ground truth, i.e., $Y = (y_1, y_2, \ldots, y_n)$, the weighted cross-entropy loss can be calculated to optimise the model,

$$l = \frac{1}{n} \cdot \sum_{i=1}^{n} - [w_2 \cdot y_i \cdot \log(p_2) + w_1 \cdot (1 - y_i) \cdot \log(p_1)] \tag{15}$$

where n denotes the number of person pairs. w_1 and w_2 denote the balanced weights of the two results of relationship existence. Suppose that there are n samples consisting of q_1 samples with $y_i = 0$ and q_2 samples with $y_2 = 1$, then $w_i = 2 \times (1 - \frac{q_i}{n})$. That balances the model learning for the imbalanced data because the person pairs with "no relation" are much more.

The non-visual cue-based component determines whether the two individuals belong to the same social group by setting the distance threshold τ. When the similarity of the trajectory pair is less than τ, it indicates that they are likely to be in the same social group. We set different values to find the optimal τ.

5. Experiment

5.1. Dataset

PLPS (Public Life in Public Space) Dataset: To comprehensively sense the attributes of citizens, Qing et al. [6] collected 71 videos from real urban public spaces. In this dataset, the individual bounding boxes of each frame are labelled so that it is easy to obtain the trajectories of all individuals by linking boxes. Besides, some other annotated attributes include social group, social relationships, human activities, etc. As for the task of social group detection, it is integrated into the baseline framework of multi-group activity recognition, so the videos are split into training/testing sets for the balanced distribution of group activities. To highlight the nature of social groups, we re-divide the data to balance the distribution of social relationships, which is also beneficial for the joint research of social group detection and social relationship recognition. After the division, 19 videos are used for testing, and the rest are for training. Due to privacy protection, the facial blur is made before release. To conduct the positional structure-based method, we followed Yoo et al. [11] to annotate individual positions and head orientations on a top view. The top-view positions are automatically computed by using depth and focal length. In this paper, the depth is estimated by the pre-trained model provided by Godard et al. [63] and the focal length is set as 7.9608. Same as [13,14], the four head orientations are annotated manually, including front, back, left and right.

Social-CAD (Social Collective Activity Dataset): To jointly learn the social groups, individual actions, and sub-group activities in videos, Ehsanpour et al. [5] annotated different social groups to extend the original CAD [28] for the study of the social group activities. The dataset is also acquired from the unconstrained real-world scenes, consisting of 44 videos. We follow [5] to use 31 videos for training and 11 videos for testing. Unlike the PLPS dataset, this dataset only annotated the bounding boxes of the keyframe, i.e., each 10th frame of all videos; hence the trajectories are relatively coarse. In the original CAD [28], individual poses, i.e., body orientations, are annotated, which can be used to conduct F-formation based methods. As for the top-view positions, we adopt the same way to obtain these above labels.

As shown in Fig. 10, we represent the classical scenarios on the PLPS dataset and Social-CAD. It can be seen that the complexity of the PLPS dataset is slightly higher because the scene scale is more significant, which can be seen from the resolution of two images, i.e., 1920×1080 on the PLPS dataset and 720×480 on Social-CAD. Besides the difference, there is a remarkable resemblance that the individuals' appearances are fuzzy. No doubt that it puts forward the challenge of feature extraction and increases the difficulty of social group detection.

Fig. 10. Scenarios on PLPS dataset (left) and Social-CAD (right).

5.2. Evaluation metrics

Same as [5], we follow the unsupervised clustering accuracy [64] as the evaluation metric of our problem. This metric introduces the Hungarian algorithm [65] to determine the optimum assignment between the detective groups and the ground truth, then sums the individuals divided into the correct social groups. Finally, the ratio of the sum to the total number of individuals is the accuracy for social group detection. It can be formulated as follows,

$$GC - Acc = \max_{m} \frac{\sum_{i=1}^{n} 1\{l_i = m(c_i)\}}{n} \tag{16}$$

where n is the total number of individuals, l_i and c_i denote the ground truth and the detection results, $m(\cdot)$ denotes all possible assignments, and $1\{\cdot\}$ is the indicator function.

5.3. Implementation details

The experiments in this paper are implemented in a deep learning framework (i.e., PyTorch) on an Nvidia GeForce RTX 2080Ti GPU. We choose the Adam algorithm [66] to optimise our model. The initial learning rate, batch size, and epochs are set as 0.01, 24, and 200. We also decay the learning rate by one-tenth of the previous per 10 epochs. In the non-visual component, the threshold τ is a hyper-parameter. Its optimal value is the six-tenths of the width of the frame.

To promote the robustness of our algorithm, we implement to augment the input data. At the stage of training, all patches were first resized into 256×256, followed by random horizontal flipping and cropping. The probability of flip is 0.5, and the cropped size is 224×224. At the testing stage, the random operations (i.e., flipping and cropping) were removed, and the input patches were resized into 224×224 directly. Image normalisation was also performed like some other tasks [67,68] for image recognition. Besides, we utilised the long sequence to extract the temporal information. Specifically, we make the keyframe the centre and find the nearest 300 frames (the PLPS dataset) and 200 frames (the Social-CAD) as one sample. We then uniformly selected three frames as inputs of the model. As for the thresholds of the similarity, both are the six-tenth width of the frame on the PLPS dataset and Social-CAD.

5.4. Results and analysis

To evaluate the performance of our algorithm, we compared it with the state-of-the-art (SOTA) methods on PLPS [6] dataset and Social-CAD [5]. Meanwhile, we also implemented the ablation study to prove the effectiveness of the sub-modules in our algorithm. Besides, the interpersonal distance is discussed to analyse the distance threshold, and the visualisation of results is represented to dissect our algorithm.

5.4.1. Introduction of SOTA methods and ablation methods

As mentioned in Section 1, few approaches are implemented in urban public spaces and most works detect F-formations for finding social groups. Therefore, besides MGAR [6] and SARF [5], we also introduce the classical methods for F-formation detection based on Hough voting scheme [13,16,17], game theory [23] and graph cuts [18]. Their details are given as follows,

HVFF [13,16,17]: Given individual positions and head orientations, these methods build accumulation spaces and find local maxima to detect F-formations based a Hough Voting scheme. They differ from accumulation strategies. In HVFF lin [13], Cristani et al. linearly accumulated the votes while Setti et al. aggregated the votes by the weighted Boltzmann entropy function in HVFF ent [16]. On top of HVFF ent, Setti et al. extended to develop a multi-scale version in HVFF ms [17].

GCFF [18]: It is an iterative method and initialises multiple F-formations. Next, a graph-cut based optimisation is used to group individuals and update the centres of the F-formations until convergence.

GTCG [23]: The authors developed a game-theoretic framework, supported by statistical modelling of the uncertainty associated with the position and orientation of people. In terms of the position and orientation, they first generate the frustum of social attention. Then, an affinity matrix is calculated by modelling the overlaps of their social attention frustums. Besides, they exploit the inter-frame smoothness between consecutive frames to face cases of noisy data.

MGAR [6]: This method proposed two indices to detect social groups, i.e., the semantic similarity of behaviour features and the spatial distance between individuals. The behaviour features are extracted by the Inception-ResNet-V2 [69], and ROI-Align [70] module and the similarities are the Cosine distance. The spatial distance is the Euclidean distance between individual bounding boxes. It should be noted that this experiment was re-conducted due to the re-division and the facial blur, mentioned in Section 5.1.

SARF [5]: This method first utilised the Inflated 3D ConvNet [71] and ROI-Align [70] module to extract the individual features. Then, the self-attention mechanism [72,73] and Graph Attention Networks [74] were introduced to refine the extracted features. Finally, the graph partition [75,76] was implemented to obtain the clusters.

Union: That is the spatial social interaction extraction in the visual cue-based component, as shown in Fig. 5. The union patch of person pair is fed into the multilevel ResNet-101 for recognising the existence of their social relationship.

Union + Seq: That is the entire visual cue-based component, including the spatial social interaction and temporal information extraction. As shown in Fig. 5, the combination of "CNN + LSTM"

is added to exploit the temporal information for relationship existence recognition.

Union + Seq + Fréchet: Our final algorithm includes the visual cue-based and non-visual cue-based components. The latter introduces the discrete Fréchet distance to measure the similarities of trajectory pairs for recognising their existence. The final results are obtained by complementarily fusing the outputs of the two components.

MFE: That is the multilevel feature extraction mechanism. To comprehensively verify its effectiveness, we conducted the compare the results with/without MFE mechanism in each ablation experiment.

5.4.2. Comparison with SOTA methods and ablation study

The experimental results on the PLPS dataset are shown in Table 1. In the table, we report the SOTA methods, the ablation experiments and our final algorithms with/without MFE mechanism. We compare the positional structure-based methods (i.e., HVFF [13,16,17], GTCG [23] and GCFF [18]) with our algorithm. It can be observed that these methods are worse than our final results, which demonstrates that those methods are not suitable for urban public spaces. That is because they only focus on positions and orientations on a top view and ignore the social context. As for the human behaviour-based method (i.e., MGAR [6]),we first compare it with our ablation algorithms, which only use the union patch of the keyframe as the input-noting that both algorithms do not involve temporal information. It can be observed that ours outperforms the baseline on PLPS by 9.66% and 11.76%, respectively. It proves that the proposed paradigm performs well in urban public spaces and hits the nature of social groups. After introducing the temporal information, the results were further improved by 4.54% and 2.53%, showing the importance of the temporal information. Finally, the visual cues are fused with the non-visual cues (the discrete Fréchet distance), and the final algorithm scores better accuracy, i.e., 78.73% and 79.05%. Besides, compared the data in columns 2 and 3, the accuracy is promoted by 2.10%, 0.09% and 0.32%. It demonstrates that the MFE mechanism can enhance the feature representations, especially without temporal information.

As shown in Table 2, additional experiments were also carried out on Social-CAD to validate our algorithm's effectiveness further. Like the experimental results on PLPS, our algorithm outperforms the positional structure-based methods, proving that ours is more robust in urban public spaces.The human behaviour-based method, i.e., SARF [5], achieves 83.0%, roughly equivalent to the accuracy of our ablation algorithm "Union" with the MFE mechanism. It is worth noting that the former introduces the temporal information with the Inflated 3D ConvNet while the latter does not. When our algorithm also adds the temporal information to promote the recognition of the social relationship existence, it outperforms the baseline on Social-CAD by 3.64% and 3.72%. These

Table 1
Experiment results on PLPS dataset, including comparison with SOTA methods and ablation study with/without MFE mechanism.

Method	GC-Acc (%)	
MGAR [6]	63.67	
HVFF ent [16]	65.37	
GCFF [18]	67.15	
HVFF ms [17]	67.35	
HVFF lin [13]	68.15	
GTCG [23]	70.98	
–	Without MFE	With MFE
Union	73.33	75.43
Union + Seq	77.87	77.96
Union + Seq + Fréchet	**78.73**	**79.05**

Table 2
Experiment results on Social-CAD, including comparison with SOTA methods and ablation study with/without MFE mechanism.

Method	GC-Acc (%)	
HVFF ent [16]	66.78	
HVFF ms [17]	69.41	
HVFF lin [13]	69.83	
GCFF [18]	71.60	
GTCG [23]	74.67	
SARF [5]	83.0	
–	Without MFE	With MFE
Union	80.63	82.46
Union + Seq	86.64	86.72
Union + Seq + Fréchet	**87.03**	**87.08**

show the high superiority of the proposed paradigm. Furthermore, introducing the discrete Fréchet distance boosts the final accuracy to 87.03% and 87.08%.

Overall, the experimental results on Social-CAD are higher than those on the PLPS dataset. It is caused by the larger scale and more complex scenarios, as shown in Fig. 10. Through comparing the algorithms with/without the MFE mechanism, we draw the same conclusion: the MFE can promote the feature representations for the blurred individuals' appearances, especially without the temporal information.

5.4.3. Interpersonal distance analysis

In the non-visual cue-based component, our algorithm calculates the discrete Fréchet distance of the trajectory pair. It sets the threshold τ for recognising the existence of a social relationship. As shown in Fig. 11, we represent the accuracy of the final algorithms using the multilevel ResNet-101 on the PLPS dataset and Social-CAD. We find that the accuracy is up to the maximum when τ is the six-tenths of the width of the frame while the widths of the two datasets are different.

There are two possible reasons, i.e., the scene types and the customs. On the PLPS dataset, the videos are all from outdoor scenarios in urban public spaces; hence people tend to stay within further interpersonal distance. However, there are partial indoor scenarios on Social-CAD where the space is narrower, so the interpersonal distance may correspondingly close. On the other hand, the video data on the two datasets are collected from different countries or regions, so people's cultural backgrounds and customs are diverse, which also causes the different interpersonal distances. These align with the proxemic findings [77,78] that interpersonal distance changes with physical constraints, culture, etc.

5.4.4. Visualisation results of positive and negative samples

To intuitively show the capacity of our algorithm, we visualise the detection results with bounding boxes in different colours in Fig. 12 (a-f). Each bounding box includes one social group or single-body individual. Our algorithm can detect the single-body individual, two-individual group, and multi-individual group even when the individuals' appearances are fuzzy and the sizes of individuals' bounding boxes are small. It proves the robustness in urban public spaces.

We also emphasise situations that are difficult to distinguish the existence of social relationships. In scenario (b), the two children in the green and blue bounding boxes appear to be the same social group from the perspective of F-formation (i.e., position and orientation) in one frame. However, they are just unfocused encounters [8,10], which can be observed from the temporal sequence in Fig. 13 (row 1). In scenario (f), the three people in the green, red, and blue bounding box are walking in the same

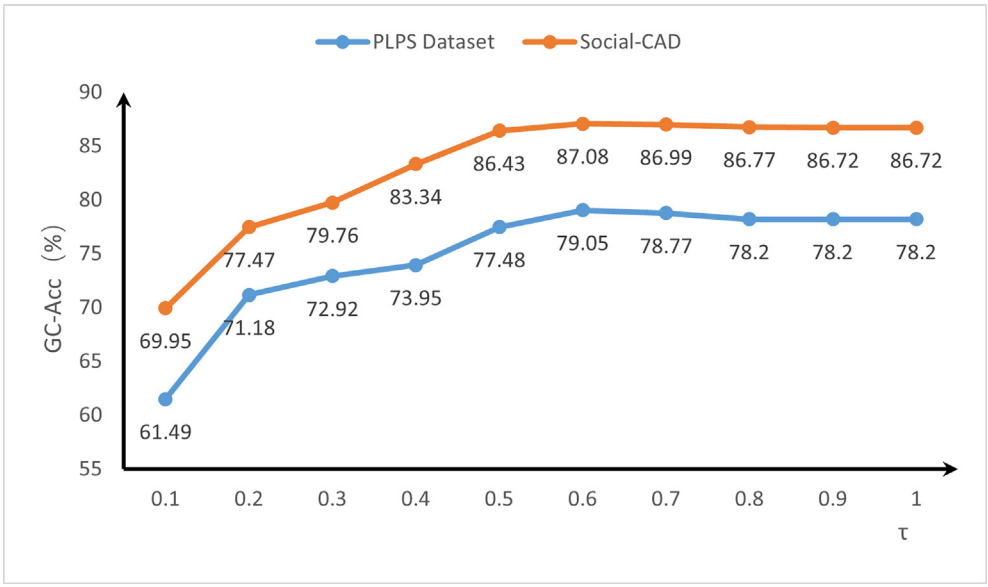

Fig. 11. Accuracy varies from the distance threshold.

Fig. 12. Result visualisation on PLPS dataset (row 1) and Social-CAD (row 2).

Fig. 13. Visualisation of the unfocused encounters on the PLPS dataset (row 1) and Social-CAD (row 2).

direction and are likely to be divided into the same social group only from the current frame. But as shown in Fig. 13 (row 2), in the temporal sequence, the individual in the red bounding box walks across the middle of the other individuals. They are also unfocused encounters, not the members of the same social group. In our opinion, the correct detection for the above scenarios is up to exploiting the spatial–temporal information. In fact, besides the spatial, social interactions and temporal information in the visual cue-based component, the discrete Fréchet distance also contains rich spatial–temporal positional information for the similarity measurement of the trajectory pairs. Besides the above positive results, we also present the negative detection to analyse the algorithm more comprehensively. As shown in Fig. 14, there are two sets of negative samples on PLPS dataset. In the blue ellipse, they are a father and daughter, while the algorithm recognises them as "no relation". It may be caused by their age gap and appearance difference, which are noise for the interaction reasoning. In addition, minor family-related samples are another important reason. For the yellow circle, the visual closeness leads to misclassification, which can be observed on Social-CAD dataset.

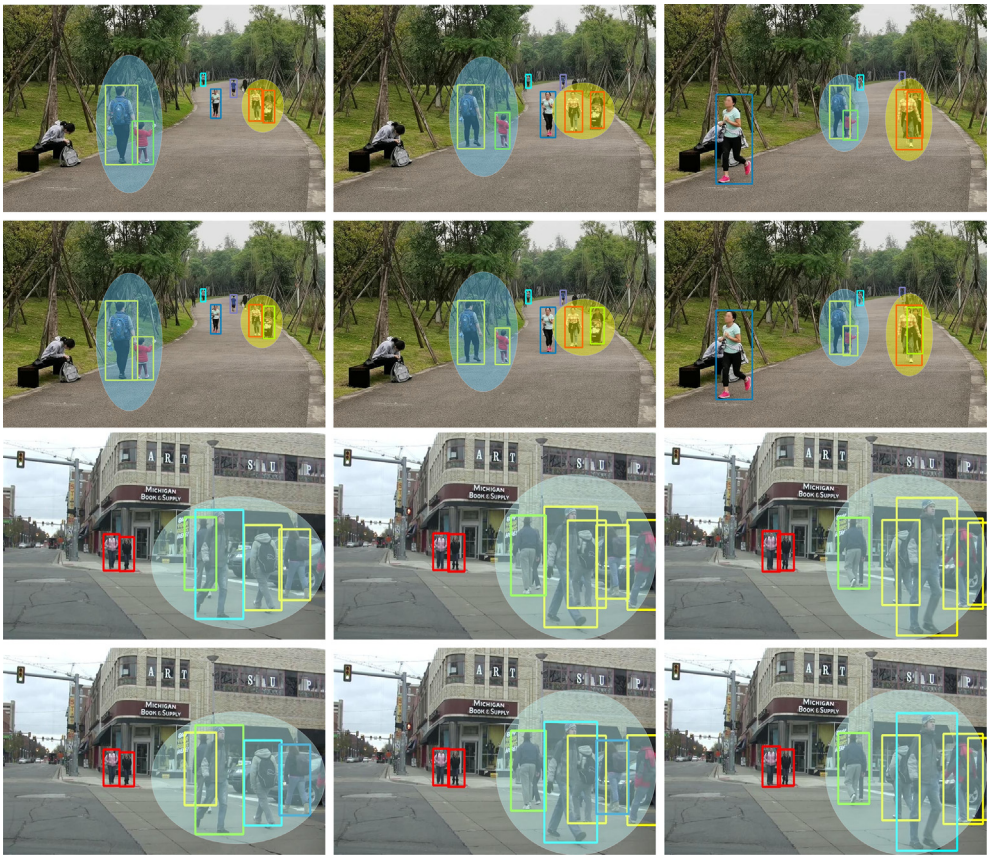

Fig. 14. Visualisation of negative detection (row 1 & row 3) and ground truth (row 2 & row 4) on PLPS and Social-CAD datasets. Individuals in the same group are labelled by bounding boxes in the same colour.

They are strangers and one individual only crosses by the other one. The possible reason is that our algorithm is not sensitive to the positional change in the depth direction. These limitations partly confine the application of the algorithm, but the current version is simple and achieves a good performance compared with the SOTA methods.

6. Conclusion

This paper summarised the existing algorithms and the inherent problem for social group detection. We proposed a new paradigm, namely relationship existence recognition-based social group detection, which can hit the nature of social groups. We also designed the corresponding algorithm, incorporating the visual and non-visual cue-based components. The former can learn spatial–temporal information through supervised deep learning, while the latter utilises the similarity of trajectory pairs to aid the existence recognition of social relationships using unsupervised index measurement. Extensive experiments were conducted to prove the proposed paradigm's superiority and the effectiveness of visual and non-visual cues. Based on the proxemics, we also discussed the differences in the interpersonal distances under the different cultural backgrounds and customs, which cannot be avoided by setting thresholds. In the future, we plan to explore the multilevel semantic context of the trajectory using the temporal networks.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Linbo Qing reports financial support was provided by National Natural Science Foundation of China.

Acknowledgement

This work was supported by the National Nature Science Foundation of China under Grant 61871278.

References

[1] J. Gehl, B. Svarre, How to Study Public Life, Island Press, 2013.
[2] J. Gehl, Life between Buildings, The Danish Architectural Press, 1971.
[3] S. Inaba, Y. Aoki, Conversational group detection based on social context using graph clustering algorithm, in: International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), 2016, pp. 526–531.
[4] H.B. Barua, P. Pramanick, C. Sarkar, T.H. Mg, Let me join you! real-time f-formation recognition by a socially aware robot, in: IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 2020, pp. 371–377.
[5] M. Ehsanpour, A. Abedin, F. Saleh, J. Shi, I. Reid, H. Rezatofighi, Joint learning of social groups, individuals action and sub-group activities in videos, in: European Conference on Computer Vision (ECCV), 2020, pp. 177–195.
[6] L. Qing, L. Li, S. Xu, Y. Huang, M. Liu, R. Jin, B. Liu, T. Niu, H. Wen, Y. Wang, X. Jiang, Y. Peng, Public life in public space (plps): A multi-task, multi-group video dataset for public life research, in: IEEE International Conference on Computer Vision Workshops (ICCVW), 2021, pp. 3611–3620.
[7] E. Goffman, Behavior in Public Places: Notes on the Social Organization of Gatherings, Free Press, 1966.

[8] T.M. Ciolek, A. Kendon, Environment and the spatial arrangement of conversational encounters, Sociol. Inquiry 50 (3–4) (1980) 237–271.
[9] S.K. Pathi, A. Kiselev, A. Loutfi, Detecting groups and estimating f-formations for social human-robot interactions, Multimodal Technol. Interact. 6 (3).
[10] H. Hung, B. Kröse, Detecting f-formations as dominant sets, in: International Conference on Multimodal Interfaces (ICMI), 2011, p. 231–238.
[11] H. Yoo, T. Eom, J. Seo, S.-I. Choi, Detection of interacting groups based on geometric and social relations between individuals in an image, Pattern Recogn. 93 (2019) 498–506.
[12] T. Yu, S.-N. Lim, K. Patwardhan, N. Krahnstoever, Monitoring, recognizing and discovering social networks, IEEE Conference on Computer Vision and Pattern Recognition 2009 (2009) 1462–1469.
[13] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, D. Tosato, A.D. Bue, G. Menegaz, V. Murno, Social interaction discovery by statistical analysis of f-formations, in: British Machine Vision Conference, 2011, pp. 23.1–23.12.
[14] L. Bazzani, M. Cristani, D. Tosato, M. Farenzena, G. Paggetti, G. Menegaz, V. Murino, Social interactions by visual focus of attention in a three-dimensional environment, Expert Syst. 30 (2) (2013) 115–127.
[15] K.N. Tran, A. Gla, I.A. Kakadiaris, S. Shah, Activity analysis in crowded environments using social cues for group discovery and human interaction modeling, Pattern Recogn. Lett. 44 (2014) 49–57.
[16] F. Setti, H. Hung, M. Cristani, Group detection in still images by f-formation modeling: A comparative study, in: 2013 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), 2013, pp. 1–4.
[17] F. Setti, O. Lanz, R. Ferrario, V. Murino, M. Cristani, Multi-scale f-formation discovery for group detection, IEEE International Conference on Image Processing 2013 (2013) 3547–3551.
[18] F. Setti, C. Russell, C. Bassetti, M. Cristani, F-formation detection: Individuating free-standing conversational groups in images, PLoS ONE 10 (9) (2015).
[19] N. Yasuda, K. Kakusho, T. Okadome, T. Funatomi, M. Iiyama, Recognizing conversation groups in an open space by estimating placement of lower bodies, in: 2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC), 2014, pp. 544–550.
[20] L. Zhang, H. Hung, On social involvement in mingling scenarios: Detecting associates of f-formations in still images, IEEE Trans. Affective Comput. 12 (1) (2021) 165–176.
[21] A. Kendon, Conducting Interaction: Patterns of Behavior in Focused Encounters, vol. 7, Cambridge University Press, 1990.
[22] T. Gan, Y. Wong, D. Zhang, M.S. Kankanhalli, Temporal encoded f-formation system for social interaction detection, in: Proceedings of the 21st ACM International Conference on Multimedia, 2013, p. 937–946.
[23] S. Vascon, E.Z. Mequanint, M. Cristani, H. Hung, M. Pelillo, V. Murino, A game-theoretic probabilistic approach for detecting conversational groups, in: Asian Conference on Computer Vision, 2015, pp. 658–675.
[24] S. Vascon, E.Z. Mequanint, M. Cristani, H. Hung, M. Pelillo, V. Murino, Detecting conversational groups in images and sequences: A robust game-theoretic approach, Comput. Vis. Image Underst. 143 (2016) 11–24.
[25] L. Cabrera-Quiros, A. Demetriou, E. Gedik, L. van der Meij, H. Hung, The matchnmingle dataset: A novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates, IEEE Trans. Affective Comput. 12 (1) (2021) 113–130.
[26] S.K. Pathi, A. Kiselev, A. Loutfi, Estimating f-formations for mobile robotic telepresence, in: International Conference on Human-Robot Interaction, 2017, p. 255–256.
[27] S.K. Pathi, A. Kristofferson, A. Kiselev, A. Loutfi, Estimating optimal placement for a robot in social group interaction, in: International Conference on Robot and Human Interactive Communication (RO-MAN), 2019, pp. 1–8.
[28] W. Choi, K. Shahid, S. Savarese, What are they doing?: Collective activity classification using spatio-temporal relationship among people, in: International Conference on Computer Vision Workshops (ICCV Workshops), 2009, pp. 1282–1289.
[29] X. Chang, P. Ren, P. Xu, Z. Li, X. Chen, A.G. Hauptmann, A comprehensive survey of scene graphs: Generation and application, IEEE Trans. Pattern Anal. Mach. Intell. (2021) 1.
[30] G. Ren, L. Ren, Y. Liao, S. Liu, B. Li, J. Han, S. Yan, Scene graph generation with hierarchical context, IEEE Trans. Neural Networks Learn. Syst. 32 (2) (2021) 909–915.
[31] T. Zhou, S. Qi, W. Wang, J. Shen, S.-C. Zhu, Cascaded parsing of human-object interaction recognition, IEEE Trans. Pattern Anal. Mach. Intell. 44 (6) (2022) 2827–2840.
[32] Y. Cheng, H. Duan, C. Wang, Z. Wang, Human-object interaction detection with depth-augmented clues, Neurocomputing 500 (2022) 978–988.
[33] S. Qi, W. Wang, B. Jia, J. Shen, S.-C. Zhu, Learning human-object interactions by graph parsing neural networks, in: European Conference on Computer Vision (ECCV), 2018, pp. 407–423.
[34] J. Li, X. Xie, Y. Cao, Q. Pan, Z. Zhao, G. Shi, Knowledge embedded gcn for skeleton-based two-person interaction recognition, Neurocomputing 444 (2021) 338–348.
[35] Y. Li, T. Guo, X. Liu, W. Luo, W. Xie, Action status based novel relative feature representations for interaction recognition, Chinese J. Electron. 31 (1) (2022) 338–348.
[36] D.B. Bugental, Acquisition of the algorithms of social life: A domain-based approach, Psychol. Bull. 126 (2) (2000) 187–219.
[37] A.P. Fiske, The four elementary forms of sociality: Framework for a unified theory of social relations, Psychol. Rev. 99 (4) (1992) 689–723.
[38] X. Chen, X. Zhu, S. Zheng, T. Zheng, F. Zhang, Semi-coupled synthesis and analysis dictionary pair learning for kinship verification, IEEE Trans. Circuits Syst. Video Technol. 31 (5) (2021) 1939–1952.
[39] L. Li, L. Qing, Y. Wang, J. Su, Y. Cheng, Y. Peng, Hf-srgr: A new hybrid feature-driven social relation graph reasoning model, Visual Comput.
[40] X. Yang, F. Xu, K. Wu, Z. Xie, Y. Sun, Gaze-aware graph convolutional network for social relation recognition, IEEE Access 9 (2021) 99398–99408.
[41] J. Gao, L. Qing, L. Li, Y. Cheng, Y. Peng, Multi-scale features based interpersonal relation recognition using higher-order graph neural network, Neurocomputing 456 (C) (2021) 243–252.
[42] S. Wu, J. Chen, T. Xu, L. Chen, L. Wu, Y. Hu, E. Chen, Linking the Characters: Video-Oriented Social Graph Generation via Hierarchical-Cumulative GCN (2021) 4716–4724.
[43] L. Zhou, J. Lv, B. Wu, Social network construction of the role relation in unstructured data based on multi-view, in: International Conference on Data Science in Cyberspace (DSC), 2017, pp. 382–388.
[44] J. Lv, B. Wu, L. Zhou, H. Wang, Storyrolenet: Social network construction of role relationship in video, IEEE Access 6 (2018) 25958–25969.
[45] L. Fan, Y. Chen, P. Wei, W. Wang, S.-C. Zhu, Inferring shared attention in social scene videos, IEEE/CVF Conference on Computer Vision and Pattern Recognition 2018 (2018) 6460–6468.
[46] L. Fan, W. Wang, S.-C. Zhu, X. Tang, S. Huang, Understanding human gaze communication by spatio-temporal graph reasoning, IEEE/CVF International Conference on Computer Vision (ICCV) 2019 (2019) 5723–5732.
[47] J. Lu, X. Zhou, Y.-P. Tan, Y. Shang, J. Zhou, Neighborhood repulsed metric learning for kinship verification, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2) (2014) 331–345.
[48] Z. Zhang, C.C.L. Ping Luo, X. Tang, From facial expression recognition to interpersonal relation prediction, Int. J. Comput. Vision 126 (2018) 550–569.
[49] Q. Sun, B. Schiele, M. Fritz, A domain based approach to social relation recognition, in: International Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 435–444.
[50] X. Liu, W. Liu, M. Zhang, J. Chen, L. Gao, C. Yan, T. Mei, Social relation recognition from videos via multi-scale spatial-temporal reasoning, in: International Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3561–3569.
[51] J. Li, Y. Wong, Q. Zhao, M.S. Kankanhalli, Dual-glance model for deciphering social relationships, in: International Conference on Computer Vision (ICCV), 2017, pp. 2669–2678.
[52] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: International Conference on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 248–255.
[53] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, A. Torralba, Places: A 10 million image database for scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. 40 (6) (2018) 1452–1464.
[54] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, A. Zisserman, The kinetics human action video dataset, arXiv.
[55] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: International Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
[56] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16×16 words: Transformers for image recognition at scale, in: International Conference on Learning Representations, 2021.
[57] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol. 1 (Long and Short Papers), 2019, pp. 4171–4186.
[58] Y. Huang, L. Qing, S. Xu, L. Wang, Y. Peng, Hybnet: A hybrid network structure for pain intensity estimation, Visual Comput.
[59] T. Liu, R. Zhao, K.-M. Lam, J. Kong, Visual-semantic graph neural network with pose-position attentive learning for group activity recognition, Neurocomputing 491 (2022) 217–231.
[60] Y. Gou, Y. Lei, L. Liu, Y. Dai, C. Shen, Y. Tong, Pretrained language encoders are natural tagging frameworks for aspect sentiment triplet extraction, arXiv.
[61] W. Li, Y. Duan, J. Lu, J. Feng, J. Zhou, Graph-based social relation reasoning, in: A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (Eds.), Computer Vision – ECCV 2020, 2020, pp. 18–34.
[62] S. Hochreiter, J. Chimidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[63] C. Godard, O.M. Aodha, M. Firman, G. Brostow, Digging into self-supervised monocular depth estimation, IEEE/CVF International Conference on Computer Vision (ICCV) 2019 (2019) 3827–3837.
[64] J. Xie, R. Girshick, A. Fahadi, Unsupervised deep embedding for clustering analysis, in: International Conference on Machine Learning (ICML), vol. 48, 2016, pp. 478–487.
[65] H.W. Kuhn, The hungarian method for the assignment problem, Naval Research Logistics Quarterly 2 (1–2) (1955) 83–97.
[66] J. Ba, D. Kingma, Adam: A method for stochastic optimisation, in: International Conference on Learning Representations (ICLR), 2015.
[67] M. Zhou, Y. Bai, W. Zhang, T. Zhao, T. Mei, Look-into-object: Self-supervised structure modeling for object recognition, in: International Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11771–11780.

[68] J. She, Y. Hu, H. Shi, J. Wang, Q. Shen, T. Mei, Dive into ambiguity: Latent distribution mining and pairwise uncertainty estimation for facial expression recognition, in: International Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 6244–6253.
[69] C. Szegedy, S. Ioffe, V. Vanhoucke, A.A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: AAAI Conference on Artificial Intelligence, 2017, p. 4278–4284.
[70] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask r-cnn, in: International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988.
[71] J. Carreira, A. Zisserman, Quo vadis, action recognition? a new model and the kinetics dataset, in: International Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4724–4733.
[72] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: International Conference on Neural Information Processing Systems, 2017, p. 6000–6010.
[73] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: International Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.
[74] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, Y. Bengio, Graph attention networks, in: International Conference on Learning Representations (ICLR), 2018.
[75] A. Ng, M. Jordan, Y. Weiss, On spectral clustering: Analysis and an algorithm, in: Advances in Neural Information Processing Systems (NIPS), vol. 14, 2001.
[76] L. Zelnik-Manor, P. Perona, Self-tuning spectral clustering, in: International Conference on Neural Information Processing Systems (NIPS), 2004, p. 1601–1608.
[77] M. Cristani, G. Paggetti, A. Vinciarelli, L. Bazzani, G. Menegaz, V. Murino, Towards computational proxemics: Inferring social relations from interpersonal distances, in: International Conference on Privacy, Security, Risk and Trust and International Conference on Social Computing, 2011, pp. 290–297.
[78] A. Sorokowska, P. Sorokowski, P. Hilpert, K. Cantarero, et al., Preferred interpersonal distances: A global comparison, J. Cross Cult. Psychol. 48 (4) (2017) 577–592.