Università
Ca'Foscari
Venezia

Master's Degree programme — Second Cycle
(D.M. 270/2004)
in Informatica — Computer Science

Final Thesis

—

Ca' Foscari
Dorsoduro 3246
30123 Venezia

An Improved Dominant-Set Algorithm for Detecting Conversational Groups in Images

Supervisor
Prof. Marcello Pelillo

Co Supervisor
Dr. Sebastiano Vascon

Candidate
Dawda Jatta

Matriculation number: 854939

Academic Year
2015/2016

ACKNOWLEDGMENT

I thank God for giving me the good health and strength to undertake this academic journey. I was able to complete this thesis due to the assistance and guidance of Prof. Marcello Pelillo as the supervisor and Dr. Sebastiano Vascon as the co-supervisor. I would therefore like to offer my cordial thanks for the opportunity to work this thesis with both of you, for your guidance, comments, correction of the thesis and encouragement. In addition, I also gained a lot from the weekly seminars at European Centre for Living Technology (ECLT) which were delivered by renowned scholars from diverse backgrounds and countries. I would therefore like to offer my sincere gratitude to Prof. Marcello Pelillo for this opportunity.

My thanks also goes to all the members of the Dipartimento di Scienze Ambientali, Informatica e Statistica (DAIS) for providing a conducive learning atmosphere in the department. The same appreciation goes to the administration of Università Ca 'Foscari Venezia and the region of Veneto for their wonderful support to the international student in terms accommodation, career guide, exchange programs with financial aid just to name a few.

Special thanks also to the local students for accommodating the international students well. The social events organized by DAIS students body did helped mingle and make new friends. Special thanks to my buddies Niccolò Turetta and Federica Armellin for the valuable information prior to my arrival and upon arrival. The same appreciation goes to my fellow international students, most notably Momodou Njie, the Ethiopians, Indians, Pakistanis,Nigerians, Cameroun etc and friends, for their excellent friendship and support during this hectic journey. Life would have been difficult without you!

I cannot end without thanking my family (especially my Aunt and Grandmother) and government of the Republic of The Gambia for their moral and financial support when it was most needed. Thank you for the trust bestowed on me.

ABSTRACT

Social encounters modelling is gaining more prominent as the need to analysis diverse social interaction groups is on the rise in both videos and images. Inspired by social psychological, social encounter attempts to analysis who is interacting with whom in a social gathering such as party, chat, etc. There exist two types of social encounters namely: focused and unfocused encounters. In this thesis, we modelled social encounter using a game-theoretical framework. We translate the notion of evolutionary game theory into social encounter where the pure strategy corresponds to the detected individuals in the scene and the payoff corresponds to similarity measure between subjects. Using this approach with position and orientation we are able to statistically modelled F-formation. We experimented this approached on several benchmark datasets and our experiment shows significant improvement in performance.

CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

In recent times visual analysis of groups are gaining more attention in many fields and applications. Traditionally, in computer vision groups analysis can be generalize in two general perspectives: meeting analysis and video-surveillance [43]. The aim is gradually shifting from encoding simple activities done by an individual subject to detecting dyads (two members) or clusters of social interaction [15,19,22,23,26,27,29,46,48,50,52]. In a broader context groups are vital in numerous areas and applications such as social and life sciences [16,41].

Fig. 1.1: A Simple Unfocused Encounter Setting

As pointed in [14] a group is a composition of two or more social entities who are related to one another. Simplifying this definition to observable properties: two or more people who are spatially and temporally close to each other with a similar velocity [3]. There exist different types of groups with differing dimension, durability just to name a few [17, 43]. Cognizant of great advances already done in single person activities [1,30,31,36,40,49], our main focus is on standing conversational groups commonly known as Facing-Formation [43, 48]. Inherited from social psychological notions and one of the most important concept used in the scheme of automated methods, F-formation [25], whose novel explanation states: an F-formation arises whenever two or more people sustain a spatial and orientational relationship in which the space between them is one to which they have equal, direct, and exclusive access. Many researches have been done by social physiologists to understand how individuals behave in public since humans are essentially social species as can be attested to

the fact that they socialize regularly to achieve their objectives. Hall [20] for instance suggested that the association and levels of interactions could be inferred by considering diverse physical distances. It is important to note that one key condition necessary for an F- formation is that the transactional segment [8], (which is the region in front of the body in which the limbs can be accessible easily) of all the members should overlap. This region can be considered as an individuals frustum of social attention.

Social encounters can be categorized into two groups: focused encounters and unfocused encounters figure 1.2 and figure 1.1. First defined by Goffman [18] focused encounters are an assembly of people who partake in a gathering in such a way that there is a common space within which a conversational exchange can occur such as freely conversing, playing a video game, watching a movie/game or discussing about a poster/art work. Thus, an F-formation can be considered as a specific instance of focused encounters which may emerge during several and different social events, such as a coffee break, party, a school/company dinner, visit in a gallery or a family fun day out in the park. In a nutshell, when people naturally decide to be in each other's immediate company to intermingle with one another. Such scenarios of real world F-formation are illustrated in figure 1.3. Different F-formation exist as illustrated in figure 1.4. Unfocused encounters on the other hand involve the way in which individuals move to ensure they avoid bumping into each other such as greeting a friend whiles passing or giving way for someone to pass.

Fig. 1.2: A Simple Focused Encounter Setting

An F- formation consist of three social spaces namely: the o-space, the p-space and the r-space. The most prominent space is the o-space figure 1.4 which is a convex empty space surrounded by the people involved in a social interaction, in which every participant looks inward, and no external people are allowed. The p-space is a narrow strip that surrounds the o-space, and it contains the bodies of the conversing people whiles the r-space is the zone outward the p-space.

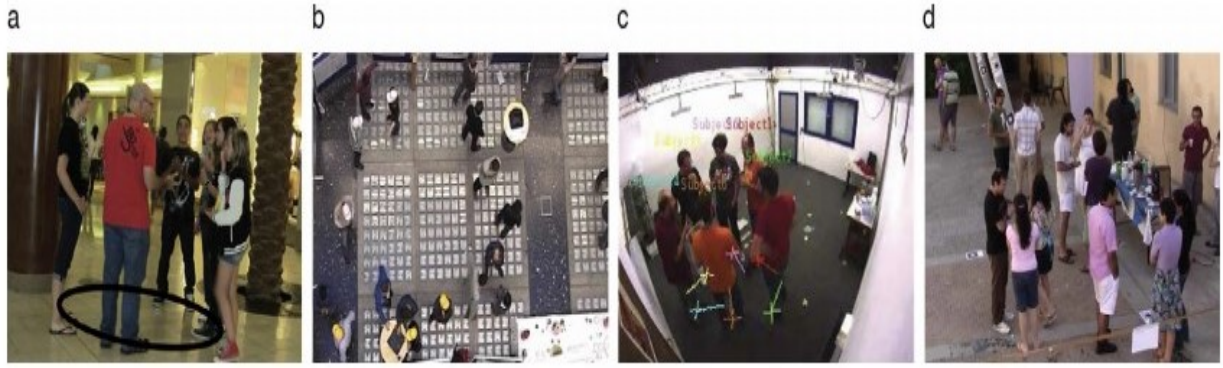

Fig. 1.3: Standing conversational groups: (a) in black, graphical depiction of overlapping space within an F-formation: the o-space; (b) a poster session in a conference, where different groupings are visible; (c) circular F-formation; (d) a typical surveillance setting where camera is located at 2.5 to 3m from the floor, for which detecting groups is challenging.

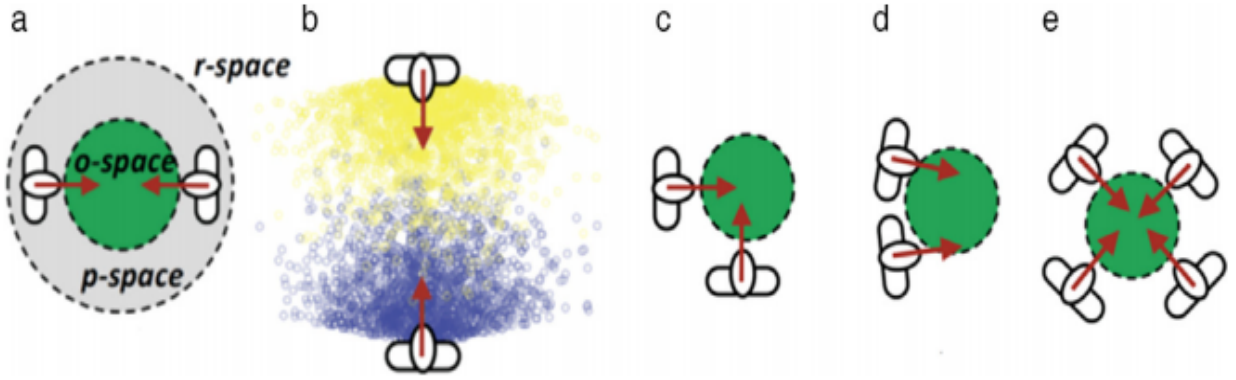

Fig. 1.4: F-formations; (a) components of an F-formation: o-space, p-space, r-space; in this case, a face-to-face F-formation is sketched; (b) modeling the frustum of attention by particles: in the intersection stays the o-space; (c) L-shape F-formation; (d) side-by-side F-formation; (e) circular F-formation.

2. MOTIVATION AND GOALS

Our approach is inspired by the work of S.Vascon et al [48] who formulated a robust approach to automatically detect F-formation using a game theoretical clustering approach. Our objective/goal is to build on this approach and make improvements to this approach by solving two key issues with this robust approach thereby improving F-Formation estimation. We should be able to automatically detect if two or more persons are facing each other or not since by theory interaction is more likely to happen if individual's frustum of social attention overlaps which is as a result of them facing each other during a conversation and also avoid particular situations that are not considered in [48]. A simple illustration in figure 2.1 highlights the two issues with the approach.

Firstly, as can be seen in (A),(B) and (C) which are queuing, using frustum overlap can be misleading, thus producing incorrect estimation. This is for the simple fact that even thought there are frustums overlapping, this does not necessarily mean the person's are in conversation since an divividual is facing the back of another individual. Because of the closeness, the frustums of the persons overlap.

Secondly, as illustrated in (D) and (E) as can be seen there are no frustums overlapping, we realized the computed similarity function Eq.4.8, always produces high a positive value because they used Jensen-Shannon (JS) divergence. The Jensen-Shannon (JS) divergence always gives high values to probability distributions that do not have anything in common. Thus, the farer the distance the higher the value. This also produces incorrect estimation. Thus, as a result of such action, the algorithm tends to group people even if they have no overlap in their frustum. These two problems served as our motivation and our goals are to find possible solutions which are discussed in Chapter 4.

Fig. 2.1: Problems of the previous approach [48]: (A),(B) and (C) are 3 persons queuing with their frustum overlapping and (D) and (E) are two person sitting side-by-side without frustum overlapping

3. STATE OF THE ART

3.1 Person Detection

In recent years, digital image collections have grown radically. As the volume of image and video data accessible increases, robust, configurable object detection systems for managing this data will become crucial. Thus, object detection systems are relevant as they are used to search through the increasing quantity of image and video databases. Automated object detection has numerous possible uses including image retrieval, surveillance applications, auto industry for designing driver assistance systems, and as front ends to recognition systems.

There are already extensive works done on object detection, but we will mention just a few relevant works on human detection. Schneiderman et al [40] proposed computer methods for automatic object detection using classifiers. In this method, they train the classifiers to form the frontal and right profile viewpoints and they built the left profile detector by reflecting the right profile detector. In order to make detection, they do a scan on each classifier over the original image and a series of resized versions of the original image. Papageorgiou et al [36] proposed a descriptive for object detection in unconstrained, cluttered scenes. Their main objective was to do a classification by means of overcomplete wavelet basis inspired by image reconstruction techniques. The general framework for this object detection approach is divided into two components: the training and testing phases. In order to do the training, the detector receives a set of images of the object class as input that have been aligned and scaled to ensure they are all in nearly the same position and the same size and secondly a set of patterns that are not in our object class. Then they performed a computation to encapsulate the vital information of each object class each of the patterns. Thus, yielding a set of positive and negative feature vectors. These feature vectors are then used in order to train a pattern classifier to differentiate between in-class and out-of-class patterns. In the final stage, that is, the testing phase, the main goal was to detect objects in out-of-sample images. The

trained classifier are used to decide which patterns show the objects of interest. They extract the same set of features just as in the training phase and then feed them into the classifier; then the classifier output determines whether or not they highlight that pattern as an in-class object. P. Viola et al [49] came up with a pedestrian detection system approach that integrates intensity information with motion information. Unlike other object detectors which naturally attempt to track objects in motion over several frames and then evaluate the motion to look for periodicity or other cues, they simply implemented a new detection style approach using information about motion as well as intensity information. They extracted the motion information in various ways, key among them optical flow and block motion estimation. The function of the block motion estimation is basically to determine the scale of the estimate which is wholly well-suited with multi-scale object detection. Gavrila et al [1] proposed a "system approach" to pedestrian detection using a system optimization scheme. This optimization scheme approach models the system as a sequence of individual components and finds a respectable overall parameter setting by merging individual ROCs by means of a convex-hull technique. In summary, they proposed a new test method for the validation of a pedestrian detection system in a real vehicle setting. Mikolajczyk et al [30] proposed a new detection method for human (Body Model) in single images which extends the detection covering full bodies as well as close-up views in the presence of clutter and occlusion. They used seven different body part for the detection namely: frontal head and face (inner frame), profile head and face (inner frame), frontal upper body, profile upper body and legs. The detection proceeds in three stages. Firstly, the features are detected individually across the image at multiple scales. Secondly, the detected individual parts are based on these features and finally bodies are detected based on assemblies of these parts. Furthermore, A. Mohan et al [31] proposed a general example-based framework that locate people in cluttered scenes. Unlike the previous method, instead of detecting the whole body this method detects the components of a person's body in an image that is the head, the right and left arms and the legs. After detecting the components it checks to make sure that the detected components are in the proper geometric configuration and then combine them using a Classifiers. The classification is done with the introduction of a new classification technique called an Adaptive Combination of Classifiers (ACC) which comprises of the distinct example-based components classifiers trained to detect different objects parts such as heads, legs and arms(right and left). Another renowned work done on person detection is Histograms of Oriented Gradients

for Human Detection [12]. This method is basically based on evaluating well-normalized local histograms of image gradient orientations in a dense grid. The rationale behind this approach was that, local object appearance and shape can often be characterized rather well by the distribution of local intensity gradients or edge directions, even without precise knowledge of the corresponding gradient or edge positions. In order to achieve this they simply divide the image window into small spatial regions or cell where each region or cell accumulates a local 1-D histogram of gradient directions or edge orientations over the pixels of the cell. Then these histograms entries are combined to form the image. I will conclude a summary on some of the works done on person detector with a brief summary of one of the most popular tools: Deformable Parts Models (DPM) detector. Felzenszwalb et al. [13] work on this model is worth noting that warrant them to receive a lifetime achievement award at the PASCAL VOC challenge. Their work have a central elements of many classification, segmentation, person's layout and action recognition tasks. The main idea behind deformable parts model was to represent an object model using a lower-resolution 'root' template, and a set of spatially flexible high-resolution 'part' templates. Each of the part captures local appearance properties of an object, and the deformations are characterized by links joining them.

3.2 Group Activity Detection

Groups are created during multi-party activities when people decide to be within each other's immediate surrounding. Groups can be described by location and movement of individuals and may vary in size and structure. In order to do grouping of groups that recognizes the activity of each group, we need to have an effective and compact group activity descriptor that can encode interactions and behaviors of people in groups. This descriptor not only should be mathematically formulated to be used in any recognition framework, but also should encode discriminative data characteristic of each group action.

There have been numerous prior works done on estimating group activities by modelling individual and group behavioral level. These prior works considered a group to be people who should be close together such as playing a video/watching a basketball game, forming a queue, having a coffee at a bar or a discussion with friends about an event. This principle is motivated by social physiological. Lan et al [26] for instance proposed three diverse methodologies to model the Group-person and person-person interaction. The motivation behind this approach was that since a group is composed of individuals, knowing the individual and group activity for instance talking, queuing, watching or facing helps to disambiguate individual human activities which are otherwise difficult to recognize. Individual action can also benefit from knowing the actions of other surrounding persons for instance two persons facing in the same direction indicates that are probably queuing or two persons facing each other indicates that they are probably talking. Yu et al. [50] also came up with a graph-cut based approach to solve the issue of discovering and analyzing interactions between individuals in social networks. They were the first to introduce the modularity-cut algorithm, that is, in the domain of computer vision. The basic notion of this modularity-cut was to dividing the social network into social groups in such a way that it minimize the connections between subgraphs, or minimize the cut size. Tran et al [46] also proposed a graph-based clustering algorithm to discover interacting groups in crowded scenes. They highlighted one key element most of the existing approaches usually treat group action recognition as a singular action done by most people visible in a scene which should not be the case especially in crowded scene. In order to discover social interaction, they used the intuition in [38] to discover socially interacting groups by discovering interacting groups as searching for dominant sets of maximally

interacting nodes in a graph which can be viewed as clustering problem using the dominant set concept. Furthermore, Mora-Colque et al. [32] developed a robust and simple approach to group people in videos. This approach has 3 main components. The first component is the segmentation component which extracts objects in motion from the scene using background removal based on the mixture of Gaussians. Then they tracked the segments extracted by optical flow and in the third and final stage they computed the relationship between segments and their labels using their flow and local features. In summary, they provided two criteria for detecting groups namely: proximity and direction flow of persons and using temporal HoG features. Ge et al. [9] developed an approach to identify a small group in crowded scenes such as people shopping, queuing to access a ticketing machine or small conversational groups using agglomerative clustering. Grouping is done by starting with an individuals as distinct clusters and they gradually built larger groups by merging two clusters with the strongest inter-group closeness using smallest Hausdorff distance. By doing this, they were able to identify small groups with high accuracy. Li et al [27] took a slightly different approach to detecting people who are sitting side-by-side such as in lecture hall or watching a poster / an art work in a gallery. They proposed a voting-based methodology based on matching against exemplars to detect and localize small-group interactions within larger social gatherings such as classrooms and lecture halls. This type of interaction is a little bit different from other as it requires a focusing of senses. Choi et al. [7] did an extensive work on modelling different group types such as queuing, standing, sitting in rows, sitting facing, sitting side-by-side and many others. The method they proposed aimed at detecting and localizing these different group structures in a single image notwithstanding their varying viewpoints, number of participants, and occlusions. They introduced the concept of discriminative interaction patterns to encode the patterns in 3D. Initially they segmented individuals in the image into diverse classes of structured groups wherein participating individuals share the same patterns of interactions. Then they localize these segmented groups in the 3D space; and finally they provided semantic descriptions to each structured group.

3.3 Group Detection Techniques using Head and Body Orientation

In recent years many works have been done that automatically detect conversational groups using positional and orientation information.

One of the earliest work done using head pose and body orientation was Bazzani et al. who [4] used a multi-target Tracker (Hybrid Joint-Separable filter) to detect person's position (head and feet location) in each of the frames and later this info is used by the head pose detector. The underlying idea concentrate on two vital social cues: visual focus of attention which helps to directs where and what a person is viewing at and it is mostly determined by head pose and eye gaze estimation and the space and environment permits to identify signals of the potential individual's interest, with respect to both the physical environment and the other members acting in the scene. They employed a clissification approach based on mutli-class boosting algorithm for effective head orientation estimation. Cristani et al. [11] proposed a statistical analysis of spatial-orientational arrangements approach that simply takes as input the positions and orientations (head) of the people in a scene; then, employing a voting strategy based on the Hough transform, it identifies F-formations and the entities connected with them. They highlighted that it is very difficult in real social scenario for individuals engaged in social interaction to orient themselves on an exact circumference and facing each other. In order to solved this they decided to inject uncertainty in their voting strategy. This proposed method used random Gaussian variables to model uncertainty of the head orientation and position and a voting stage where each samples cast a vote for a candidate position which are then accumulated in an intensity accumulation space. Positions with strong weights votes receives high votes. Setti et al. [42] also proposed a multi- scale extension of the Hough-based approach [11] by having different F-formation modeling cardinalities. They envisaged an F-formation involving more than 2 person arranged in a circle, their locations modelled as the vertices of a regular polygon which lies in the interior of the personal image. By doing so they were able to define the radius associated with an F-formation with different cardinalities and collect a set of radii which they exploited for individual grouping. The same approach used to estimate head pose and orientation in [11] is used. Hung and Kröse [23] detected F-formations by formulating the problem in terms of identifying dominant set , a form of maximal cliques in weighted graphs via graph-theoretic clustering. Using the definition of F-formation and the bases of criteria of a clusters,

that is, internal and external criteria, they were able to identify members of an F-formation and their associates thinking about clusters as dominant set using socially motivated estimate of focus orientation (body orientation) to determine body orientation. The underlying idea was that, people in a scene are mostly likely to orient their body towards the person(s) they are motivated to be in conversation. Setti et al. [47] did a comparison analysis of [11,23] highlighting their strengths and weakness. They noted that using position and orientation [11] performs better whiles [23] outperforms [11] using only position. Tran et al. [45] followed an approached of the graph based [23], extending it to deal with video- sequences and identifying several categories of group events such as walking, Waiting, Queuing, Walking, Talking, Dancing and Jogging. Using the technique in [38] to discovering interacting groups as a graph based clustering problem using the dominant set concept, they began by first searching for dominant set in the graph, then by eliminating that set of vertices from the graph, and iteratively repeating this procedure with the remaining set of vertices, until there remain no dominant sets in the graph. The leftover vertices after the elimination of found dominant sets represents individuals who are not associated with any group. This method is similar to [23] but a little bit different in the way the weights of the graph edges are calculated for instance it exploit social cues to compute the weigh and it approximates an individual attention as an ellipse centered at a fixed offset in front of him. Thus, the more the ellipses overlap the more interaction occur. S.Vascon et al [48] considers an F-formation as a non-cooperative clustering game exploiting the approach proposed in [2] in which they modelled the uncertainty associated with position and orientation by statistical means. Furthermore, they employed the notion of multi-payoff evolutionary game theory to integrate temporal information over multiple frames. Similar to [48] but did not dealt with conversational groups, Choi et al [7] proposed a method for identifying and localizing groups by demonstrating the different forms of group behavior discriminatively by encoding the relationship between individuals in 3D. Both approaches are similar in the sense that they find overlaps in a sample space. More recently, Setti et al. [43] proposed the Graph-Cuts for F-formation (GCFF) method which is based on a graph-cuts frame-work for clustering individuals in still images. The basic notion of GCFF was it finds the o-space of F-formation irrespective of its arrangement and assignments it to individuals who transaction segments overlap. All of these methods discussed above are based on one of the appraoches in [34]

4. OUR CONTRIBUTION AND APPROACH

4.1 Our Contributions

Our contributions in this work are two folds: that is, to solve the two issues highlighted in Chapter 2. A summary of our final pipeline that solved the issues raised in Chapter 2, is summarized in figure 4.4.

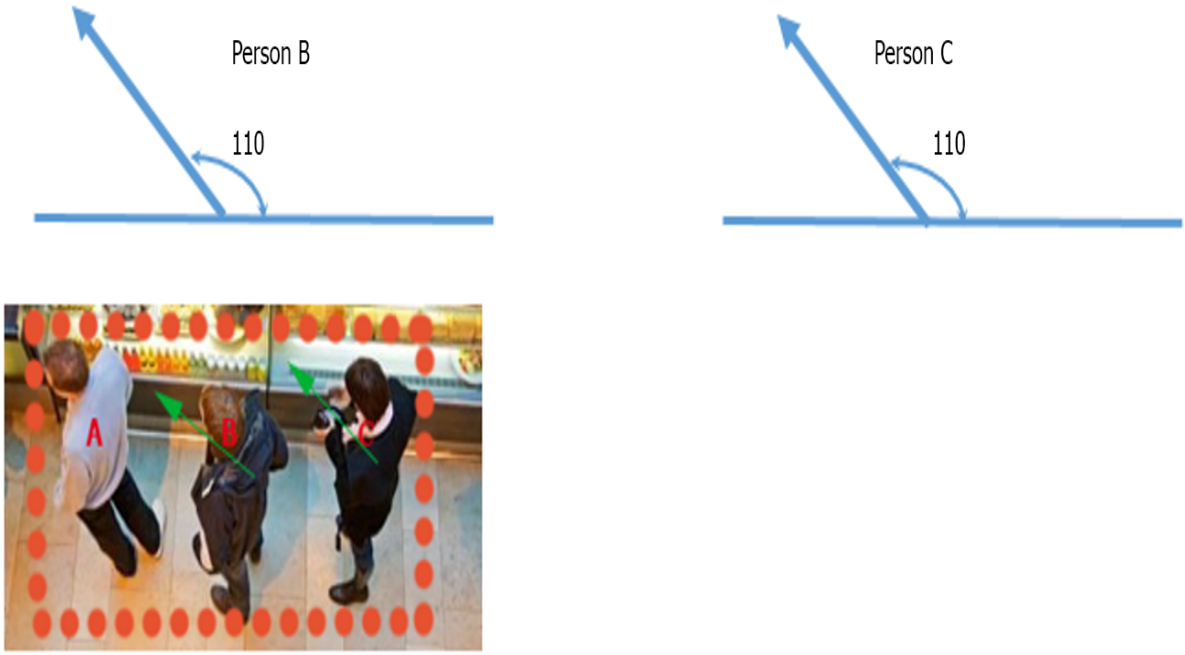

Fig. 4.1: Simple example to demonstrate how to determine if two person are not facing

In our first contribution, given a person's position and body/head orientation, we determine if two or more persons are facing or not. This is simply done by introducing a function that checks if two or more individuals are facing or not according to their angles using the cosine similarity in Eq.4.1. This is then represented in a simple binary matrix where 1 indicates facing, 0 otherwise.

Basically, if two individuals have the same angle with an overlapping frustrum, it simply tells us they are queuing or facing the same direction. This is then represented in a binary matrix in which 0 means not facing and 1 otherwise.

$$
\begin{aligned}
x_1 = cos(\theta_1) \quad y_1 = sin(\theta_1) \\
x_2 = cos(\theta_2) \quad y_2 = sin(\theta_2) \\
z = x_1 * x_2 + y_1 * y_2 \\
CF_{i,j} = \begin{cases} 1 \ \ if \ \ z < 0 \\ 0 \ \ otherwise \end{cases}
\end{aligned}
\tag{4.1}
$$

Where θ_1 = angle of person 1 and θ_2 = angle of person 2

Using a simple example as in figure 4.1, where we simple take Person B and C in our queue sample picture, where we assume the two individuals are looking alomost in the same direction (110 degrees). We can determine if they are facing or not according to our check facing function using the cosine similarity. Below is how we go about it:

$$
\begin{aligned}
z = cos(110)^2 + sin(110)^2 \\
z = 0.1170 + 0.8830 \\
z = 1 \Rightarrow CF = 1
\end{aligned}
\tag{4.2}
$$

Using Eq.4.1, from our result for the value of z in solution 4.2, we can conclude that B and C are not facing since the value is positive.

Furthermore, we use another scenario where two persons are facing figure 4.2, where Person A and B are sitting facing each other at 300 and 110 degrees respectively. Thus, using our check facing function in Eq.4.1:

$$
\begin{aligned}
z = cos(300) * cos(110) + sin(300) * sin(110) \\
z = 0.5000 * (-0.3420) + (-0.8660) + 0.9397 \\
z = -0.1710 - 0.8138 \\
z = -0.9848 \Rightarrow CF = 0
\end{aligned}
\tag{4.3}
$$

Thus, from the solution in 4.3, we could see that our function value is less than 0. Hence, in our check facing matrix, we simply put a 1 indicating the two persons are facing.

Fig. 4.2: Simple example to demonstrate how to determine if two person are facing

Our second contribution was to solve the second issue raised in Chapter 2, that is, the function in Eq.4.8, always produces positive value even if there is/are no frustrum overlapping which should not be the case with no overlapping given higher value and overlapping low value. Thus, the approach is unsuitable for the simple fact that it assigns low value to frustum overlapping. In order to solve this, we redefined the similarity function in Eq.4.8 that utilizes frustum overlap in the similarity function. Using assumption from sociological point of view, two or more individuals whose frustum do not overlap are not supposed to be interacting. Thus, we simply set the output value Eq.4.8 to zero since there is no frustrum overlap, otherwise we just take the actual function value. By doing this, we forcing the ESS- Clustering to divide persons that are not facially interacting.The redefined function is in Eq.4.9. The soultions given above to the problems of the previous work [48] help to improve the detection of scenarios which were not considered like detecting queues.

Now that we have our frustum overlap (FO) and check facing (CF) of person i and j, we then combine the two matrices into one, A, by simply doing a point-wise multiplication of the two as shown 4.4:

$$A_{i,j} = FO_{i,j} * CF_{i,j} \tag{4.4}$$

By doing so, our check frustum overlap acts as a filter by removing all the incorrect detection of the frustum overlap. By doing this combination, we were able to solved the problem of queues which involve frustum overlapping and thus would give us a better F-Formation estimation.

4.2 Our Approach

Our approach takes as input set of frames containing the person's position and head/body orientation as in [48]. These two components can be easily obtained nowadays even if the estimation is not done accurately. It is noted that several methods [6, 24, 29] are aimed at extracting such information from raw images/video sequences. The following steps and figure 4.3 below summarize the pipeline of the old approach and our new approach in figure 4.4:

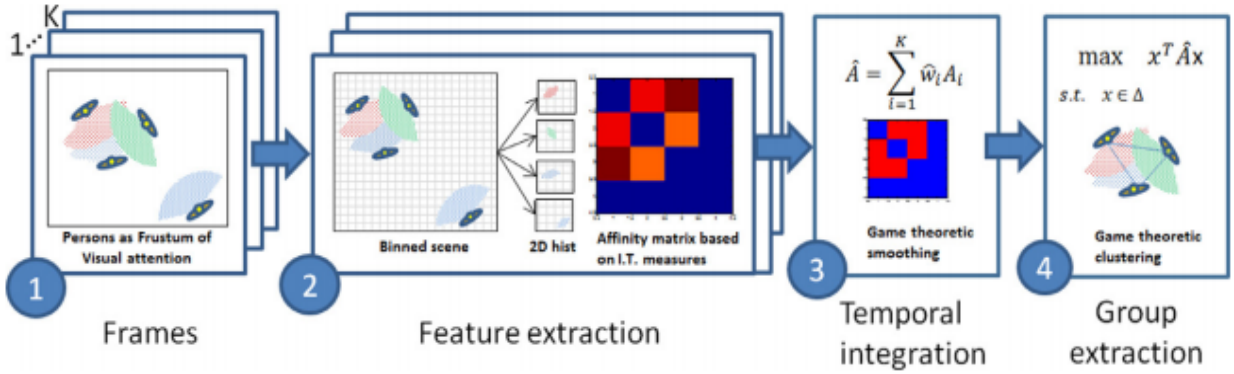

Fig. 4.3: Previous algorithm's Steps

1. Based on each person's position and body/head orientation in real world coordinates, $p_i \in P$ in every frame, we generate a frustum f_i and modelled by a 2D Histogram.

2. Then do a computation of the pairwise affinity matrix for each , $p_i \in P$.

3. In case of a smoothing across multiple frames (which we are not interested in our case), compute the weights of each frame based on the theory of multiple payoff games.

4. Finally, we extract the F-formations using evolutionary stable strategy (ESS)-Clusters.

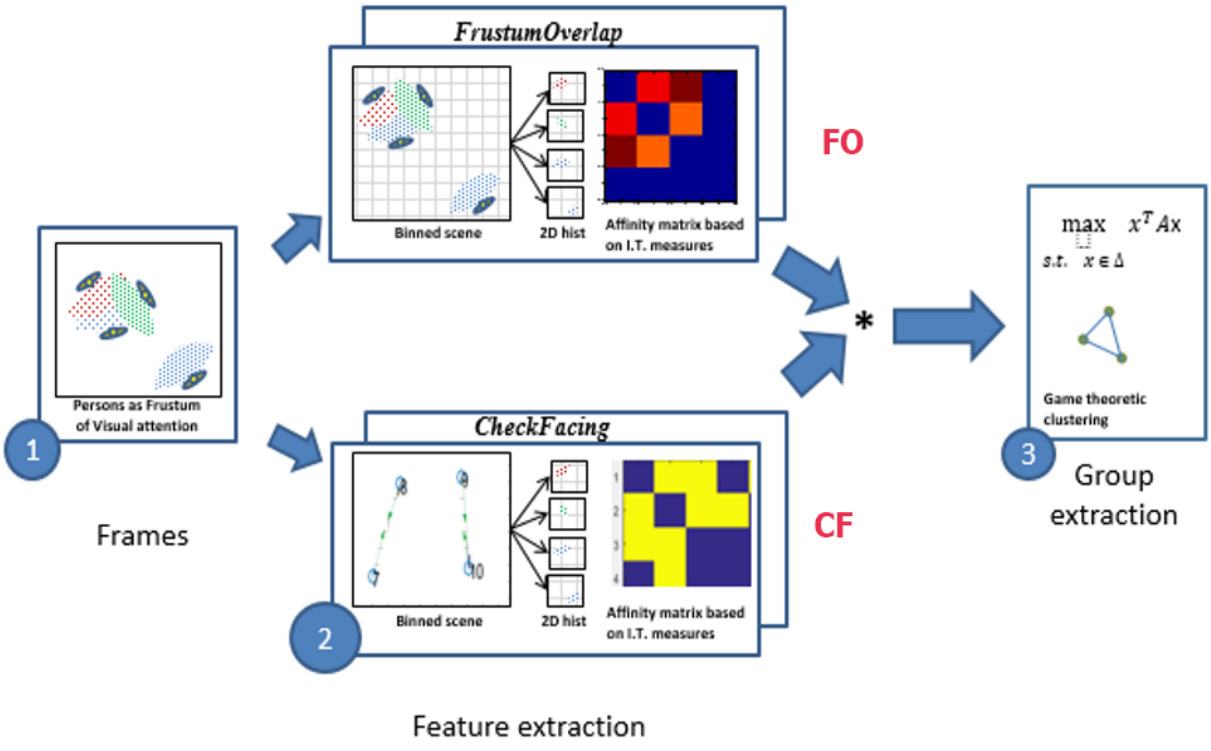

Fig. 4.4: New Algorithm's Steps

4.3 Modelling frustrum of attention

In this work we will used the new frustum model proposed in [48] figure 4.5b which is based on two probability distribution sampling. Their frustum of social attention was inspired by Kendon's definition of transaction segment. Kendon's transaction segment takes in consideration two vital components:*field of view of the individual and the locus of attention of all the other senses given body orientation*. In normally scenario it is much easier to get the head pose rather than body orientation in crowded settings, thus the head pose helps provide an approximation of the direction of the social attention of the frustum.

Three elements play a vital part in determining the social attention of the frustum of a given individual namely: a direction θ which is the individual's head orientation, an aperture γ=160° which is used to determine the range of possible eye gaze direction given head pose and finally the length l in *cm* or *meters* depending on a given data.

The new proposed sampling approach using the G *distribution* and B *distribution* has two advantages compared to the old one figure 4.5a in the sense that it decouples the whole algorithm making use of samples and histograms which makes the entire approach non-parametric and thus easily integrable with upcoming models. Moreover, since we are dealing with unpredictable data such as detections, head orientation etc which are prone to errors and noisy, data smoothing is necessary.

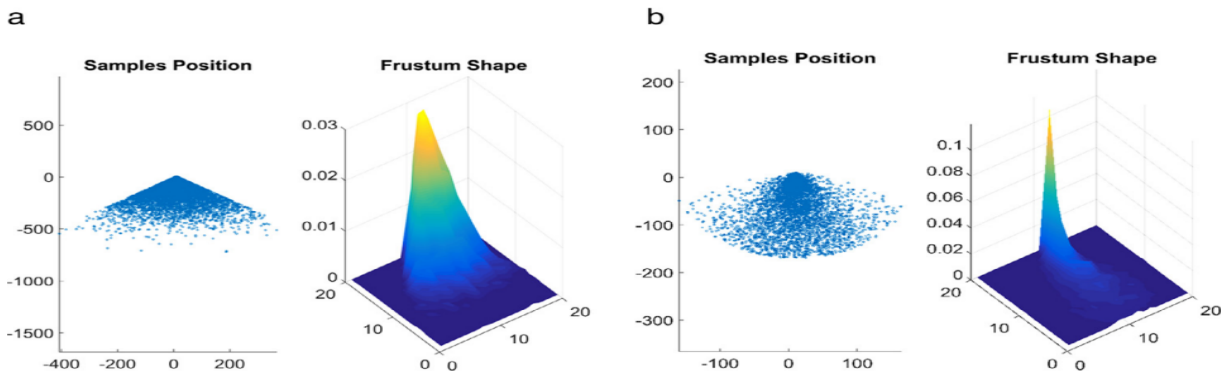

Fig. 4.5: (a) The old frustum model (b) The new frustum based on the G and B distribution.

The sampling technique used in the previous were taken from a 2D Gaussian distribution which is not ideal for the field of view as it is chopped figure 4.6a. This is because sample are labelled either valid or invalid and the annotating only ends when the required number of valid samples are reached. This procedure makes the approach time consuming but with the new proposed sampling only the valid samples are considered making the process faster and more efficient. Also the new method is more descriptive of the decaying nature of the human view. From the above the new *sampling method is thus different from the old one in the sampling method and peripheral view*.

The function of the G *distribution* figure 4.6a in this approach is basically to generate samples associated to the frustum's aperture. This is positioned in the head orientationθ of the individual with a variance set in such a way that the entire width of the Gaussian distribution tallies to the anticipated frustum's aperture. As can be noted in the Gaussian, 99 % of the samples are situated

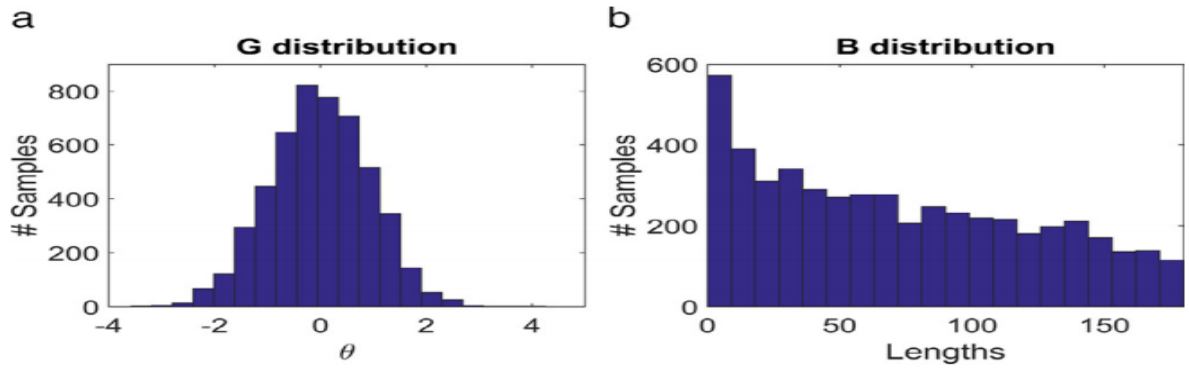

Fig. 4.6: (a) The probability distribution over the orientations and (b) the distance from the person.

in the interval of $[-3\ \sigma,\ 3\sigma]$ this interval will match to the full frustum's aperture so that setting the variance in such a way that the aperture is completely enclosed becomes an easy assignment, $\sigma = \frac{1}{3} * \frac{\gamma}{3}$

The B *distribution* figure 4.6b function is to provide samples that are dense in close proximity of the individual while it decays as you move a little far away. In order to accomplish the required shape as used in [11] we set the parameter of the distribution $\alpha = 0.\ 8$ and $\beta = 1.1$. The values of the B distribution are in the interval [0, 1] and these values are then multiplied by the required frustum's length l. the final samples acquired from these two distribution are in angles and distances and in order to obtain the 2D they applied a simple trigonometric on each of the samples. For instance assumed a pair of sample from both distribution (G_i, B_i) with position of an individual (p_x, p_y), for each sample, the 2D position is:

$$\begin{aligned} s_x &= p_x + Cos(G_i) * B_i * l \\ s_x &= p_x + Sin(G_i) * B_i * l \end{aligned} \tag{4.5}$$

Using the above equations, we can acquire *n* a set of samples that falls in the human frustum of visual attention given *n* independent samples from both G and B*distribution*.

4.4 Histogram binning

In order to decide the best binning for our 2D histogram, we used the same procedure as in [48] where they carried out an extensive experimentation on all the publicly available datasets. In our

case we narrow down our search from [5-200] with a step of 5 and obtain the best with the highest F-score. This is because we already know the best bins in [48] and thus if there is going differences if would not be that much. In doing so we noticed that bins 20 and 80 gives the highest F-scores for all the publicly available datasets. We decided to stick with bin 80 since it gives the highest F-score except for FriendsMeet2 dataset where we stocked to the old parameters that is bin 20.

4.5 Quantifying pairwise interaction

In order to measure interaction we used the concept of frustum overlap that is two individuals are expected to be interaction if their social attention frustum overlaps. Moreover, to encode the the uncertainty about the true transactional segment of the individuals given head pose, we measure the pairwise interaction as a distance between distribution. We consider information theoretic measure in modelling the distance between since we are dealing with histograms that represents probability distribution.

Assume that a pair of discrete probability distribution $P = \{p_1, ..., p_n\}$ and $Q = \{q_1, ..., q_n\}$ we used the intuition of Kullback-Leibler (KL) divergence, a well-known concept in probability theory and information theory which is a way of measuring the distance between two probability distribution. This is defined as:

$$D(P||Q) = \sum_{i=1}^{n} \log p_i \frac{p_i}{q_i} \tag{4.6}$$

This class of divergence is known to be asymmetric but there exist an alternative symmetric version to Kullback-Leibler divergence known as Jensen-Shannon (JS) divergence [28] which is defined as:

$$J(P, Q) = \frac{D(P||M) + D(Q||M)}{2} \tag{4.7}$$

Where M is the average of the two distribution ($M = \frac{1}{2}(P + Q)$). Thus given two individuals *i* and *j* in a scene and there vectorized histograms h_i and h_j, we can calculate the distance between *i*

and *j* as either $D(h_i||h_j)$ or as $JS(h_i, h_j)$. Furthermore, in order to acquire the measure of affinity rather than distance between each pair of histograms, we utilized the classical Gaussian Kernel:

$$\gamma(i,j) = \exp\{-\frac{d(h_i, h_j}{\sigma}\} \tag{4.8}$$

Where *d* = either the KL- divergence or JS-divergence. One important parameter in the above equation is the parameter σ which helps determine how distant individuals are from each other in an F-Formation. This parameter and the length of the frustum l are determine in each dataset using σ ranging from [0.1 – 1] with a step of 0.1 and *l* ranging from [50-200] with a step of 5. Finally, once the measure of affinity is computed using the above equation then it becomes possible to detect groups composing of individuals who are possibly interacting by exploiting a grouping game using evolutionary game theory.

$$\gamma(i,j) = \begin{cases} \exp\{-\frac{d(h_i,h_j}{\sigma}\} & if \quad i \bigcap j \neq 0 \\ 0 \quad otherwise. \end{cases} \tag{4.9}$$

5. THE DOMINANT SETS CLUSTERING BACKGROUND

First introduced by Pelillo and Pavan [38] in 2003, dominant set framework is renowned for graph-theoretic notion of a cluster which simply generalizes the idea of maximal clique to edge-weighted graphs. This approach has a number of numerous applications [37, 38, 48, 51] and has proven to be fast and efficient. Furthermore, there have been a number of generalization of these approach to hypergraphs and multigraphs recently in [5, 33]. This has also demonstrated to be a fast and effective method in modelling the structure of the underlying data manifold.

The main goal of clustering is to partition a set of objects into groups in which objects in the same group have the same mutual similarities and objects outside of the group have a greater dissimilarities to the inside objects.

Clustering application can be found in many areas such as medical imaging, Bioinformatics, signal processing, market research, social network analysis, image segmentation, educational data mining, crime analysis just to name a few. The task of clustering can be divided into two: Central (featured based) and Pairwise (Graph based).

In our case, we will mainly focus on the graph-based approach. This method represents the input objects as feature vectors, thus each object is represented as a point in an n multidimensional space in which the similarities and dissimilarities can be calculated using different clustering measurement techniques such as Euclidean distance or Manhattan. Despite its numerous applicability, there exist several cases where its representation as feature vectors cannot be easily determined. Thus, in such case the clustering objects are represented in graphs and the similarities between graphs have to establish using one the many techniques.

Computing the connection between arbitrary graphs is NP-hard but some instance of graphs can be computed in polynomial time. Since we can calculate the similarities and dissimilarities between objects which is not possible using feature based clustering then the most suitable method is pairwise clustering. The basic intuition of this clustering approach is that given the similarity

matrix of objects as an input, then based on some coherency criteria it tries to create a partition on a given data. This is done irrespective of how the data is represented and therefore it is consider as a more appreciative since it considers all form of representation.

In the Dominant set framework pairwise clustering mechanism is used where the data to be clustered are represented as an undirected edge-weighted graph with no self-loops. For instance given n objects that are denoted by vertices, and the edge which join them is labeled with the similarity value (weight) between the joined objects. Finally, the graph is then represented as an *n* by *n* similarity matrix in which the value of the matrix are the weights that defines the corresponding similarity of the points of the corresponding column and row. As an example, given matrix with*M* and the similarity between vertex *i* and j is represented in $M_{i,j}$. Bearing in mind that our original graph has no self-loops which simply means no link is connected to itself so our similarity matrix will have zero diagonals. Let's consider a simple scenario in which we have a binary matrix (0, 1) corresponds to either similar or dissimilar without taking into consideration their degree of similarity. As in such the graph is an undirected and unweighted graph. This property exhibited by such graph satisfied both the internal and external criteria of the classic concept of graph theory which is the view of maximal clique figure 5.1 .

Cliques are formed when there exist a mutually adjacent relationship between vertices in a graph. Let *D* be a graph, a clique in G is a complete subgraph of *D*. Let *T* be a subset of *V* where *V* is a vertex in G, such that every two vertices in T are connected by an edge in G. There exist different forms of cliques: *Maximum Clique, Maximal clique, and Strictly Maximal Clique*. A clique is termed *maximum clique* if it has the biggest cardinality. When a clique is not contained in any larger clique or not part of a larger clique it is called *Maximal clique*. A *Maximum Clique* is always maximal clique but the contrary cannot hold since is not a subset of the bigger clique and cannot be extended by adding an adjacent vertex. Due to the unstable nature of maximal cliques figure 5.2, a Strictly Maximal Clique is a stable set which must fulfill the following conditions: if a graph is maximal clique, all the vertices outside it can't have a number of edges, incident on its vertices, which is more than one less the cardinality of itself. As stated earlier, due to the fact that maximal cliques are unstable in some cases, when one of the vertex is dropped from the maximal clique and another vertex added from outside, a new maximal clique will be formed.

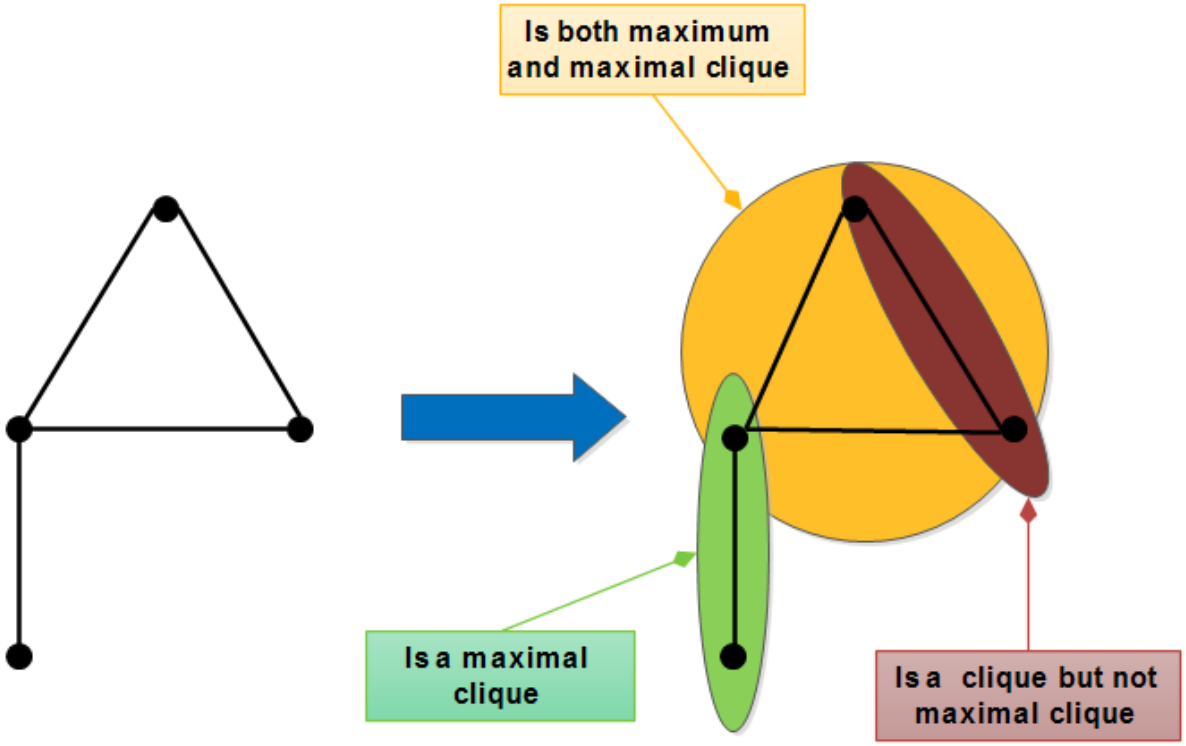

Fig. 5.1: Clique,maximal clique and maximum clique example

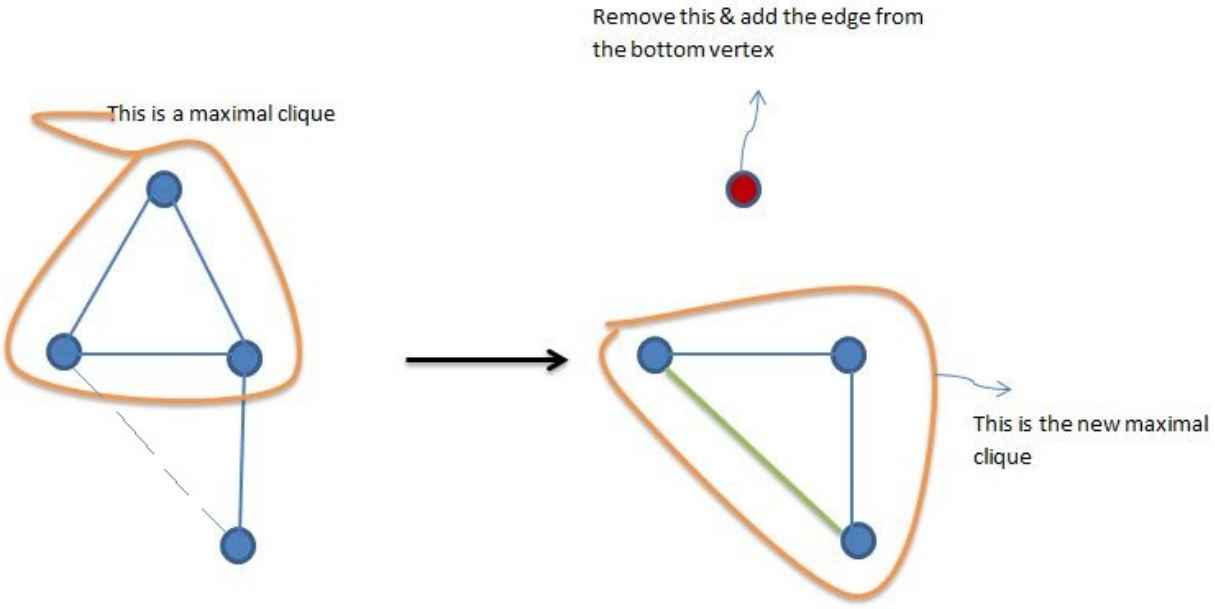

Fig. 5.2: Instablity of maximal clique

The perception of cluster concurs with the concept of maximal clique for instance in the case when we have a binary similarity. However, this is not the case with an edge weighted graph. This is where the concept of dominant set framework emanates in order to generalize the maximal clique view into an edge weighted graph. The dominant set framework is a combinatorial concept in graph theory with the primary aim to generalize the idea of maximal clique to an edge weighted graph. Prior to giving a formal definition as defined by Pelillo and Pavan [38], lets deal with some basic definition of dominant set.

Definition: This can be define as the sum of the weights, that is the similarities of the edges that connects one point to the rest of the points in the set divided by the cardinality of the set gives us the average weighted degree of a point(AWDegs). To formulate this mathematically, the average weighted degree of a point i figure 5.3 with respect to a set of vertices S that is not an empty set is expressed as:

$$AWDegs_(i) = \frac{1}{|S|}\sum_{j\in S} w_{i,j}$$

Where $w_{i,j}$ represent the weight between the two vertex *i* and *j*

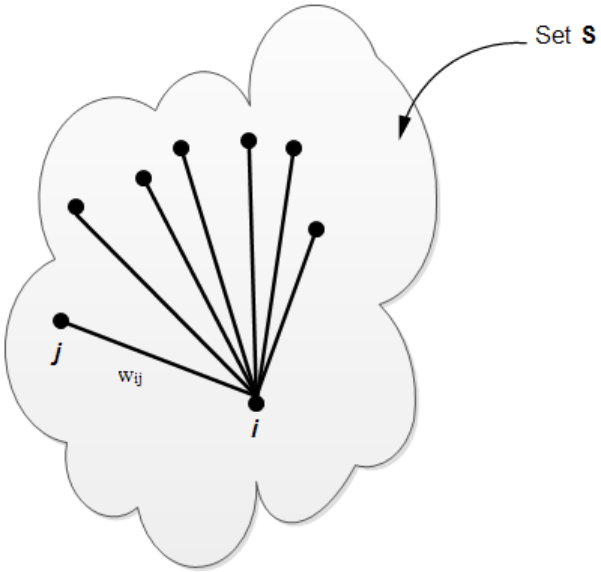

Fig. 5.3: Average weighted degree of point *i*

Here the relative similarity,$\phi_s(i,j)$, between two points (vertex),*i* and *j* figure 5.4, where *j* is not

an element of the set *s*, it measures the similarity between node *i* and *j* with respect to the average similarity between nodes *i* and its neighbors in set *S*.

$\phi_s(i,j) = w_{(i,j)} - AWDegs(i)$

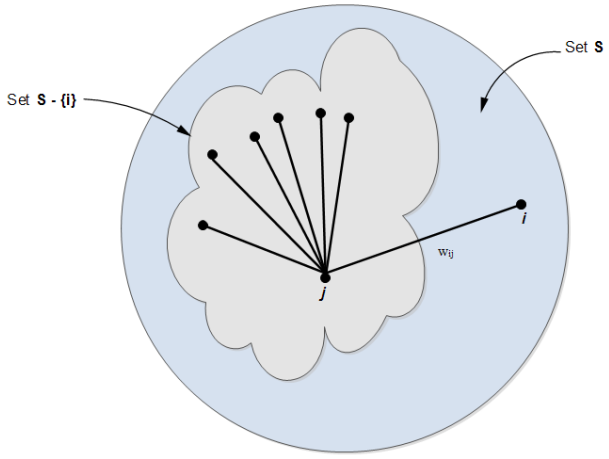

Fig. 5.4: Relative similarity between two objects *i,j*

As such the value of $\phi_s(i,j)$ could fluctuates between negative value and positive depending on the value of the absolute similarity. If the absolute similarity is less than the average weighted similarity, the value will become negative and the contrary also holds.

Therefore, if the cardinality of the set *S* is 1 then by definition WS(i) = 1, that is the weight of*i*i with respect to *S* if not we have to add all the relative similarities between *i* and the rest of the nodes in the set *S*, and with this we can see how closely related is point *i* with respect to the rest of the points in *S*.

$$W_S(i) = \sum_{j \in S \setminus \{i\}} \phi_{S\setminus\{i\}}(j,i) W_{S\setminus\{i\}}(j)$$

In summary we can compute the weight of the set *S* by adding up each weights WS(i) at this point. As we noted earlier $WS_(i)$ evaluates how closely the vertex is related with other sets of nodes.

Example: Given the below graph figure 5.5, lets calculate its weights.

Solution: We first need to find the vertices with the smallest and largest weight. We simply do this by summing the edge weights of the links which are directly connected to that node. For

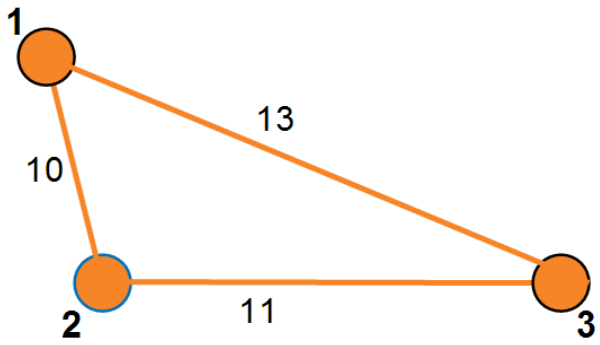

Fig. 5.5: Example on weighted graphs

instance summation of edge weights of node 1 is 10 + 13 = 23, node 2 is 10 + 11 = 21 and node 3 is 11 + 13 = 24. We notice that $W_{\{1,2,3\}}(3) > W_{\{1,2,3\}}(1) > W_{\{1,2,3\}}(2)$.

Now lets use the formal formulas and do a comparison with the above results. But lets bear in mind the following:

$W_{\{i,j\}}(i) = W_{\{i,j\}}(j) = \omega_{i,j}$

$W_i(i) = 1$ (by definition when the cardinality of the set S is 1)

$\phi S_{i,j}(j) = \omega_{i,j}$

$W1;2;3(1) = \phi_{\{2;3\}}(2;1)W_{\{2;3\}}(2) + \phi_{\{2;3\}}(3;1)W_{\{2;3\}}(3)$

$= \omega_{2,1}AWDeg_{\{2,3\}}(2)\omega_{2,3} + \omega_{3,1}AWDeg_{\{2,3\}}(3)\omega_{2,3}$

$= (1011/2)11 + (1311/2)11$

$= 132$

Repeating the same procedure for $W_{\{1;2;3\}}(2)$ and $W_{\{1;2;3\}}(3)$, we got 104 and 140 respectively. Now we need to do a summation of all the 3 results to get the total weight of S, W(S) which is 132 + 104 + 140 = 376.

Now let do more experiment on this graph and see what the effect would by adding external vertices to this existing graph. Then later we would check the sign of the vertices including the newly added vertex which shows the global similarity of this new vertex with respect to the previous vertices. Lets say we add vertex 4 and 5 to our initial graph figure 5.6.

Calculating the weight for the newly added vertices, we notice that vertex 4 has a weight

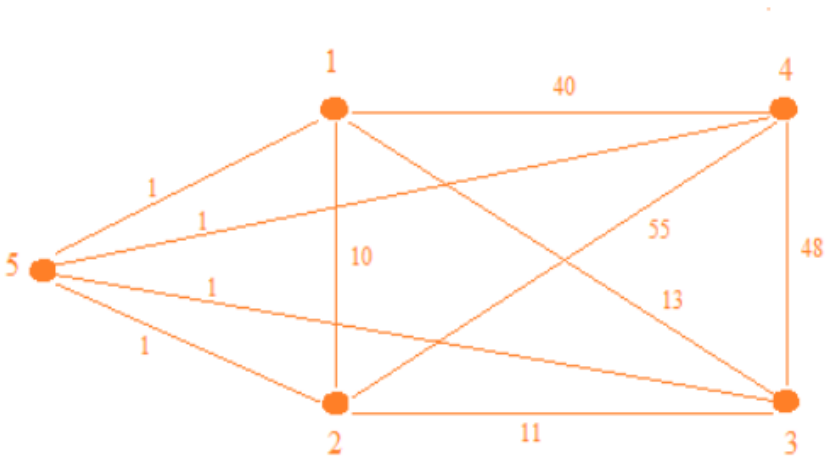

Fig. 5.6: New Node 4 and 5 added to the previous Graph

$W_{\{1,2,3,4\}}(4) > 0$ and vertex 5 $W_{\{1,2,3,5\}}(5) < 0$. Thus we can conclude that vertex 4 is highly similar/bounded/grouped to 1, 2, and 3 since the weight between them is high. on the contrary, node 5 is loosely bonded with the existing graph since $W_{\{1,2,3,5\}}(5) < 0$ (that is negative total weight). Thus, we can conclude by saying that adding vertex 5 to the graph will weaken/decrease the overall weight or coherency property of the graph.

Definition, Pelillo and Pavan [38]: A non-empty subset of vertices $S \subseteq V$ such that $W(T) > 0$ for any non-empty $T \subseteq S$, is said to be dominant if:

1. $W_S(i) > 0$ for all $i \in S$
2. $W_{S\cup\{i\}}(i) < 0$ for all $i \notin S$

The above properties satisfies the clustering criteria, that is cohesion and separation. In conclusion we can bravely say that the two notions of clustering and dominant set coincides. But now given a data how can we partition it into dominant set? Instead of using a standard algorithm to find dominant set, Pelillo and Pavan [38] transformed the purely combinatorial problem of finding a dominant set in a graph into a pure quadratic optimization problem and use evolutionary game theory dynamical system to solve the optimization problem.

5.1 Dominant Sets and the Quadratic Optimization

Now that we have the basic idea about dominant set, let deal with its mathematical representation in details. The mathematical representation will allow us to partition the data set in clusters in terms of dominant set framework. As stated earlier, finding maximum clique is NP but nonetheless we can still find it in quadratic time. As we have stated previously, Pelillo and Pavan [38] transformed the purely combinatorial problem of finding a dominant set in a graph into a pure quadratic optimization problem and use evolutionary game theory dynamical system to solve the optimization problem. This kind of problem is a well-known general form from graph theory, Motzkin-Straus problem [33].

Motzkin-Straus theorem : given an undirected unweighted graph *G = (V, E)* and W is the adjacency matrix of the graph. There is a one-to-one correspondence between the clique number of the graph $\omega(G)$ and the maximal optimizer of the problem:

$$\begin{aligned} &maximize \qquad \mathbf{f}(\mathbf{x}) = \mathbf{x^T W x} \\ &\text{subject to } x \in \Delta \\ &\textit{If } \mathbf{x}^* \textit{ is the maximizer, then } \omega(G) = \tfrac{1}{1-f(x^*)} \end{aligned}$$

The standard simplex Δ is a simple geometrical structure which satisfies these criteria's, all the values of x_i should be either greater than or equal to 0 and if we sum up all x_i we should get 1. Let's see the pictorial representation of standard simplex with n equals to two and three figure 5.7.

Scholars have differing approaches to solving clustering problems for instances S. Sarkar and K. L. Boyer [39] used a slightly different but similar quadratic program to the above mentioned. They solved the spectral clustering problem by finding the biggest eigenvalue and the associated eigenvector of the given similarity matrix W. They achieved this by maximizing this quadratic problem:

$$maximize \quad \mathbf{x^T W x},$$

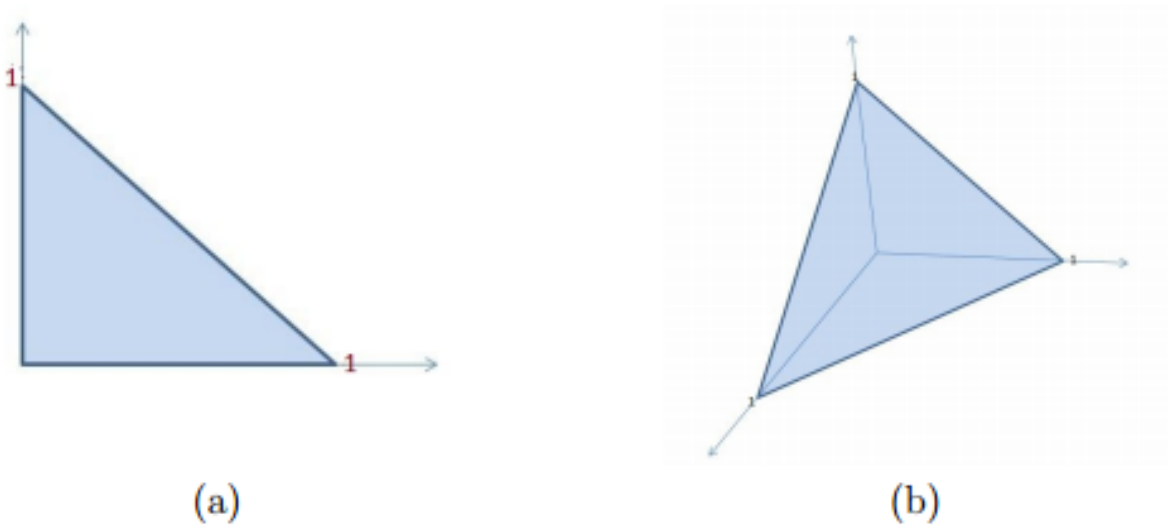

Fig. 5.7: (a) and (b) represent standard simplex where n=2 whereas n=3 respectively

$$\textit{subject to } x \in S(theSphere) \textit{ (that is } x^T x = 1\textit{)}$$

We can clearly noticed that there is a similarity in main objective function but the difference lies in the domain. This technique has some drawback most notable is that it only considers the positive eigenvalues in addition to this finding the max value is NP hard since it is simulated to the maximum clique. This drawback is considered as a limitation of this technique. Before going into detail to establish the theorem 1-to-1 correspondence between dominant sets and the strict local maximizer of the quadratic program $x^T W x$ over the standard simplex, let's deal one definition.

Definition: *A weighted characteristic vector is a vector in the standard simplex that can be defined as follows:*

$$x^S = \begin{cases} \frac{W_S(i)}{W(S)} & if\ i \in S; \\ 0 & otherwise. \end{cases}$$

Theorem, Pelillo and Pavan: *If S is a dominant subset of vertices, then its weighted characteristic vector x^S is a strict local solution of objective function in the standard simplex.*

Conversely if x^* is a strict local maximizer of the objective function in the standard simplex then its support

$$\sigma = \sigma(x^*) = \{i \in V : x_i^* \neq 0\}$$

is a dominant set provided that $w_{\sigma \cup \{i\}} \neq 0 \quad \forall i \neq \sigma$

The elements of the characteristic vector is equal to the number of vertices that exist in a graph in which their i^{th} value will equal to 0 whenever $i \neq S$, if not as mentioned previously in the theorem the ratio is taken. By doing a summation of all the elements of the characteristic vector will lead us to the same point in the standard simplex. Therefore, since the ratio of all the component vector are non-negative the weights are positive and the sum of the 3 components equal to one. In a scenario where S is a dominant set the vector is the strict local maximizer of $x^T W x$ in the standard simplex. This implies that if we are given that x^* is the strict local maximizer of the above objective function in standard simplex, by bearing in mind only the support of the vector which relate to the subset of nodes in a given graph which correspond to dominant set. We realized that the dominant sets are characterized as a continuous form optimization problem and we need to find a way to solve the optimization problem and obtain the dominant sets. One simple and effective solution to solve this optimization problem to extract dominant set from a graph is the use replicator dynamics. This is a method developed and studied in evolutionary game theory and is defined as:

$$x_i^{(t+1)} = x_i^{(t)} \frac{(Ax^{(t)})_i}{(x^{(t)})' A (x^{(t)})} \tag{5.1}$$

For all *i = 1 ... n* gives us the discrete-time version of the first-order replicator equations. In the standard simplex $\triangle$ all trajectories will remain in the simplex $\triangle$ for all future times. Moreover, when considering symmetric matrix A we can proof our objective function $f(X) = X^T A X$ will increase strictly along any non-constant trajectory of the above formula (formula for replicator dynamics for discrete time) its asymptotically stable points are in one-to-one correspondence to strict local solutions, in turn, correspond to dominant sets for the similarity matrix A.

Definition: *A Nash equilibrium is an Evolutionary Stable Strategy(ESS) if for all strategies y*

$y^T Ax = x^T Ax$ *implies* $x^T Ay > y^T Ay$.

Here assuming both x and y are Nash equilibrium then we will end up with the same result when y is playing against its opponent and when x plays with himself, by changing the role of y and x we end up $x^T Ay > y^T Ay$. If strategy y is equally good strategy as x and if we change the role of them and if my opponent choose to play y, it is advantageous if i play x against y rather than playing strategy y. Hence the above conditions provides Nash equilibrium the resistance to any small permutations in the components of vector and allows it to be more stable. So the notion of Evolutionary Stable Strategy is the notion that we are searching for our purpose of clustering. From the above we notice that the evolutionary game theory satisfied both the internal and external criteria for clustering. In our optimization problem, in a doubly symmetric game that is A=AT, we noticed that both notions of Nash equilibrium and Evolutionary Stable Strategies coincides in the standard simplex $\triangle$. As Nash Equilibrium is local maximizer of $x^T Ax$, the Evolutionary Stable Strategies is the strict local maximizer of $x^T Ax$. Since both Evolutionary Stable Strategy and dominant set coincide, we can give identical definition for both symmetric and non-symmetric scenarios of dominant set.

Theorem,Pelillo & Pavan: *A point $x \in \Delta$ is the limit of a trajectory of the replicator dynamics starting from the interior of the standard simplex iff x is a Nash equilibrium. Further, if point $x \in \Delta$ is an ESS then it is asymptotically stable.*

6. INTRODUCTION TO EVOLUTIONARY GAME THEORY

Charles Robert Darwin, FRS (Born 12 February 1809 –19 April 1882) was an English naturalist and geologist, best known for his contributions to evolutionary theory is referred as the father of evolution. In 1859 he published his book *"On the Origin of Species"*, in which he overwhelming rejected earlier the scientific concepts of transmutation of species. Darwin originated the phrase *"survival of the fittest"*. Darwin noted that in every environment, organisms need resources to survive but the resources are limited to satisfy every organism. Thus, this leads to *competition* for the available resources. He noted that there are *variations* within any group of organisms and due to this variations certain organisms may have better competitive advantages over others for scarce resources. With overtime, organisms with lesser advantages for competing to get the resources will fail in their quest to acquiring the necessary resources to be healthy and to survive. For instance if this reason for being lesser advantageous is due to genetically inherited variation, organisms with the less beneficial genes will be extinct before passing their genes to the next generation. Likewise, organisms with better competing advantages will continue to survive, probably reproduce and pass on their genes on to the next generation. Thus, this means that only the better advantageous organisms will be represented in the generation.

It was not until in the 1870s that much of the scientific community and the broad public began to accept his evolution concept as a fact. From the 1930s to the 1950s many modern evolutionary synthesis emerged in which a broader consensus was established in which natural selection was the basic mechanism of evolution. The basic idea of natural selection is the pruning out of organisms with genes that are less fit. In summary, this Darwin's concept is the unifying theory of the life sciences, clarifying the diversity of life.

A Game can be considered as an interaction between several elements/decision-makers called *players* in which each of these elements/decision-makers has a set of decisions on how to behave. These are called *strategies* and each of these strategies is associated with a *payoff* that

depends on the strategies selected by the opponents. The payoff is simply a gain or loss to each elements/decision-makers from the decisions they made. A Game can be play simultaneously and independently if not the game should define the order in which the players select their strategies. Formulating the problem into Game theoretical terms (Game theory) it is the mathematics of how elements/decision-makers (players) bear in mind conflicts of interest and prospects for cooperation when making decisions about how to behave in a Game with the opponents. It is assumed that every elements/decision-makers knows everything about the structure of the game and that the opponents fully understand the game too. One key assumption made in Game theory is that each player is concerned in maximizing their own benefit.

As an example of Game theory application, let's consider a game involving two Athletes. These Athletes are considering whether to cheat using drug for performance-enhancement or not to cheat. Our decision-maker (players) are the Athletes, the strategies are to cheat or not to cheat. There are four possible outcomes for this game:

1. Athlete A decides to cheat and Athlete B also decides to cheat.

2. A decides to cheat and Athlete B decides not to cheat.

3. A decides not to cheat and Athlete B decides to cheat and

4. A decides not to cheat and Athletes B also decides not to cheat.

They must make the decisions simultaneously without consulting each other. But in order to make a decision the Athletes need to know the payoff associated with each outcome. The best possible payoff for cheating is 4. If an Athlete cheats and his opponent does not cheat, he will loss and gets 0 for that matter. If both of them decide to cheat their payoff is 3 and if both of them decide to not to cheat they will have a payoff of 2. We formulate this in a simple payoff matrix Table 6.1.

Now which choice should Athlete A and B choose from the possible outcomes that would maximize their payoff? Let's do an analysis of each of the Athlete and see the possible outcomes. Let's start with Athlete A. Athlete A may think of a decision that would maximize his payoff in the case Athlete B decides to cheat. Thus, if Athlete B cheats, Athlete A would be better off cheating too since 3 is greater than 0. How about if Athlete B decides not to cheat? Athlete A might decides to cheat because it still maximizes his payoff. Thus, whatever Athlete B does, Athlete A is always

Tab. 6.1: Simple Game

(A, B) Payoff		**Athlete B**	
		Cheat	**Don't Cheat**
Athlete A	**Cheat**	3,3	4,0
	Don't Cheat	0,4	2,2

better off cheating. On the other hand, Athlete B might consider the following options bearing in mind the possible choices Athlete A might make. If Athlete A cheats, he might also decides to cheat. If Athlete A decides not to cheat, Athlete B will still be better off cheating too because that maximizes his payoff. One vital thing noted here is that whatever option Athlete B chooses, Athlete A is always better off cheating. Likewise whatever option Athlete A chooses, Athlete B is always better off cheating. This is called a *dominant strategy* for both Athlete A and B. A player should always play a dominant strategy when it is available irrespective if what the opponent does.

Another famous example Game theory application is the Prisoner's Dilemma Table 6.2 involving two players. The assumption is that two suspects were apprehended Police and were being interrogated simultaneously in different room. Since there is not enough evidence to convict them, the Police decided to make an offer to each suspect. The offers were either to betray or cooperate. If either suspects betrays his partner he will receive a less sentence and his partner will received a great sentence. If both cooperates they will each have a sentence of 1 year. Anyone who betrays and the partner cooperates he will be set free and the partner will be convicted for 10 years. And if both betrays they will each receive 5 years jail term. Both will be seeking an advantage over the other during the interrogation.

Tab. 6.2: Prisoner's Dilemma

(x,y) payoff	Cooperate	Betray
Cooperate	**(-1,-1)**	**(-10,0)**
Betray	**(0,-10)**	**(-5,-5)**

Doing the same analysis as in the first example, both suspect might come to the conclusion of something like: “if my partner decides to cooperate, I will be better off betraying since 0 is better than -1.”If my partner decides to betray, I am still better off betraying since -5 is better than -10”. Therefore, we say both suspects have a dominant strategy of betray and this corresponds to Nash equilibrium. This a simple and powerful concept proposed in [35] by John Nash in 1950 about behavioral reasoning in games. The main idea is that even when there are no dominant strategy players are expected to adapt strategies that are best responses to the other. If a players adapt strategies that are best response to each other, no player has the incentive to unilaterally deviate from it.

6.1 Formulating F-Formation as a Non-Cooperative Game

In order to do the grouping, we follow the intuition proposed in [44] in which the problem of detecting an F-Formation is formulated as a non-cooperative clustering game. As highlighted in [48] choosing such an approach for detecting an F-Formation has a number of advantages which includes:

1. The required similarity function does not have to be metric, thus making it usable with the Kullback-Leibler.

2. There is no need for setting a-prior number of clusters as in other clustering methods like k-means since the number of groups is anonymous.

3. Using such an approach especially when dealing with differing temporal instants for instance in multiple frame in video sequences provide us with theoretical foundation to integrate multiple payoff metrics.

Formulating an F-Formation as a grouping game, we envisage a scenario where a set of elements $O = \{1, \ldots, n\}$ and an *n x n* matrix $A = (a_{ij})$ of (possibly asymmetric) affinities which simply quantifies the similarity between the elements in *O*. Two players who are aware of the set up play a game by simultaneously selecting elements from *O*. Each player then receives a payoff proportional to the affinity that the chosen element has with respect to the element chosen by the opponent after showing their choices. Formalizing the above in game-theoretic jargon the set of elements $O = \{1, \ldots, n\}$ are the *"pure strategies"* accessible to each players and the affinity matrix A represents the *"payoff" function*. To be precise, a_{ij} represents the payoff received by a player playing strategy *i* against an opponent player playing strategy *j*. in an F-Formation set up the elements to be grouped that is the *"pure strategies"* are basically the individuals detected in the social encounter and the payoff function is the similarity measure between subjects.

Furthermore, a key notion in game theory is that of a *mixed strategy*, a probability distribution $x = (x_1, ..., x_n)^T$ over the available pure strategies *O*. Mixed strategies are constrained to lie in the (n-1) dimensional standard simplex:

$\Delta = \{x \in \mathbb{R}\} : \sum_{i=1}^{n} x_i = 1$ and $x_1 \geq 0, i, ..., n$

Where a mixed strategy x $\in \Delta$, the support of the set of elements chosen with non-zero probability can be define as:

$\sigma(x) = \{i \in O | x_i > 0\}$

Thus, the expected payoff received by an individual playing mixed strategy y against an opponent playing mixed strategy *x* is given by: $y^T Ax$. The best replies against mixed strategy *x* is the set of mixed strategies is: $\beta(x) = \{y \in \Delta : y^T Ax = max_z z^T Ax\}$. Finally, a strategy *x* is said to be a *Nash equilibrium* if it is the best reply to itself, that is if $x \in \beta(x)$ or if $x^T Ax \geq y^T Ax$ for all $y \in \Delta$. At *Nash equilibrium* no player have an incentive to switch strategy unilaterally.

All along, it is assumed that this clustering game is played within an evolutionary game setting in which two players are chosen randomly from the large population sample to play a pre-assigned

strategy. Contrary to old-fashioned game setting, given a mixed strategy $x \in \Delta, x_j (j \in O)$. is expected to represent the proportion of players that is programmed to select pure strategy *j*. Therefore, in our case of grouping, we assumed that a dynamic evolutionary process will occur that will make selection from state *x* and will evolve according to survival-of-the-fittest, that is the better-than-average pure strategies will survive while those with lower-than-average will be extinct. As such, a mixed strategy $x \in \Delta$ is said to be *an evolutionary stable strategy (ESS)* if it is a Nash equilibrium and if, for each best reply evolutionary stable strategy (ESS) if it is a Nash equilibrium and if, for each best reply *y* to *x*, we have $x^T Ay > y^T Ay$. This condition ensures that any small deviation from the stable strategies will lead to a negative payoff (doesn't payoff).

As expressed in [5, 44], ESS as a clustering game incorporates two fundamental properties of *internal coherency* and *external incoherency* of a cluster:

1. **internal coherency:** objects within the same cluster should have the same mutual similarities;

2. **external incoherency:** as the number of external objects introduced increases, the total cluster internal coherency decreases.

One key feature of this approach is it generality as it allows for more scenarios when one is dealing with a variety of integrated framework. In a scenario where the affinity *n x n* matrix $A = (a_{i,j})$ is symmetric, that is, $A = A^T$, the concept of *an evolutionary stable strategy* cluster corresponds with that of a dominant set [38], which sums to finding a local maximizer of $x^T Ax$ over the standard simplex Δ.

7. EXPERIMENT AND RESULTS

To evaluate our methods, we experimented our approach on numerous publicly available datasets Table 7.1 for F-Formation estimation in which each single frame is analyzed independently. The results were compared with state-of-the-art.

7.1 Dataset

Tab. 7.1: Datasets info

DataSet	**Number Of Sequences**	**Consecutive Frames**	**Automated Tracking**
FriendsMeet2	15	10,685	NO
CoffeeBreak	2	45,74	YES
CocktailParty	1	320	YES
Gdet	5	132,115,79,17,60	YES
PosterData	82	1	NO
Synth	10	10	NO

We used the seven current publicly available benchmark dataset Table 7.1 for F-Formation detection in which the person's x,y head orientation and position in the scene are provided. The table provides info on the number of sequences, number of frames, consecutive frames and automated tracking for each of the dataset.

FriendsMeet2 (FM2) proposed in [39] is composed of 53 sequences of which 15 are original real sequences in which the head orientation is manually annotated in each of the sequence totaling

10,685 frames. There are 16,286 frames for all the 53 sequences. This is so far the biggest group detection dataset publicly available to date. The annotation for the head orientation was done in the image plane by pointing to the head of the individual in the scene and drawing a line in the direction where the person is looking. Through the available homography has been possible to convert the line from image plane to world coordinates obtaining the real head angle on the ground plane. The ground-truth for the groups is the same as in the original dataset.

PosterData [23] is a recording of about 3h of a scientific meeting in a large atrium of a Hotel. It consist of 50 individuals involved in poster presentation and a coffee break. A number of frames totaling 82 were selected with the intuition of maximizing variance and crowdedness of the scenes Fig.7.1. A total of 21 annotators were grouped into 3-person subgroups (7 groups) and they were asked to identify F-formations and their associates from static images. A person's position and body orientation was manually labelled and recorded as pixel values in the image plane.

Fig. 7.1: Sample Frames of the poster frames Dataset.

Cocktail Party [52] dataset contains about 16 minutes of video recordings of a cocktail party in a 30 m2 lab environment involving 7 subjects Fig.7.2 . The event was recorded using four synchronized angled-view cameras (15 Hz, 1024 × 768 px, jpeg) installed in the corners of the room. Subject's positions and horizontal head orientations were logged using a particle filter-based body tracker with head pose estimation. Since this dataset was the first where proxemic information is estimated automatically, errors may be present. Groups were annotated manually by a trained expert every 3 s, resulting in a total of 320 distinct frames for evaluation.

CoffeeBreak [22] consist of maximum of 14 individuals organized in groups of 2 or 3 people

Fig. 7.2: Sample Frames of the CocktailParty Dataset.

each and focuses on a coffee-break setting of a social event. This was captured from a single camera with resolution of 1440 × 1080px. The positions of the people have been estimated by exploiting multi-object tracking on the heads, and head detection has been performed afterwards in which only 4 possible angles were considered (front, back, left, right). Later the tracked head orientations and positions were projected onto the ground plane. A psychologist annotated the videos indicating the groups present in the scenes for two sequences with combine 120 frames (45 frames for Seq1and 75 frames for Seq2) Fig.7.3 .

Fig. 7.3: Sample Frames of the CoffeeBreak Dataset.

Synth [22] generated by a psychologist for 10 different cases of F-Formation and each repeated 10 times with slight variation in person's head orientation and position resulting in 100 frames. There are 3 groups with an average of 9 individuals in the scene and some of the individuals do not belong to any group. In this dataset, there is noiseless in in position and head orientation. Fig.7.4 shows examples of frames taking from the dataset.

GDet [22] was captured around a vending machine where area where people were taking cof-

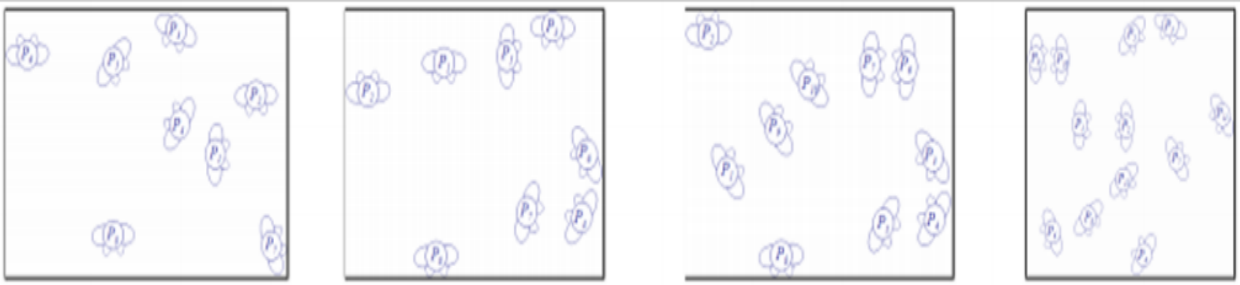

Fig. 7.4: Sample Frames of the Synthetic Dataset.

fee, drinking and chatting. It consist of 5 subsequences of images attained by 2 angled-vie with low resolution cameras of 352 × 328px with a number of frames ranging from 17 to 132, for a total of 403 annotated frames Fig.7.5. The head orientation is only limited to 4 possible angles.

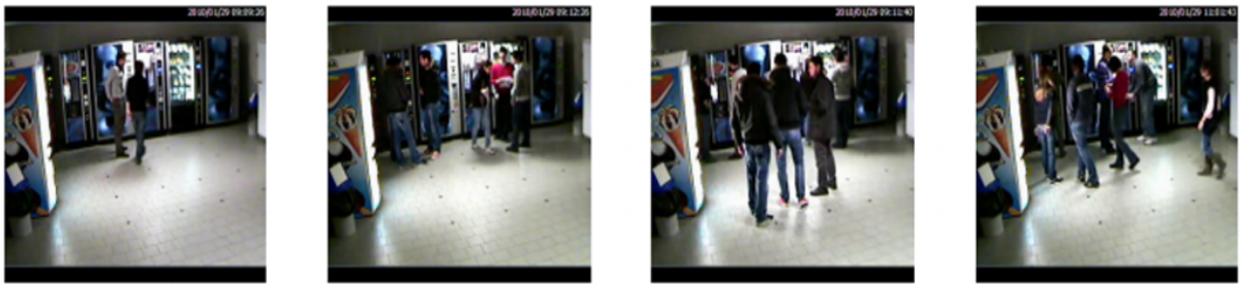

Fig. 7.5: Sample Frames of the GDet Dataset.

7.2 Evaluation metrics and parameter exploration

In order to do the evaluation, we used the measure of accuracy as in [48] in which a group is correctly estimated if at least $[(T \cdot |G|)]$ are correctly detected by the algorithm and if no more than $[(1 - T) \cdot |G|]$ false subjects are found where $T \in [0, 1]$, knowns as tolerance threshold and |G| is the cardinality of the labelled group G. In our case we used two parameters for T, that is $T = \frac{2}{3}$ and T = 1. For each of the frame the correctly detected groups: true positives-TP), the miss-detected groups: false negatives-FN) and the hallucinated groups: false positives – FP were determined. Using this evaluation metrics, we compute the precision, recall, and F1-score (harmonic mean of precision and recall) per frame, then the average over all the frames is used as the final scores:

$$\textbf{recall} = \frac{TP}{TP+FN}, \textbf{precision} = \frac{TP}{TP+FP}, \textbf{F1} = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Using similarity function in Eq.4.4 with dissimilarity function in Eq.4.7 as used in [48], we explored and validated different parameter combinations to examine the performance. This was done by varying the parameter σ in the range [0.1–0.9]. We also did an exploration on the effect of the frustum length and our analysis is based on the finding in [10, 21] in which the proximity of focused encounter between individuals is likely to take place is taken to be between *45 cm* to *2 m*.

The results in Table 7.2 shows results of our approach compared with R-GTCG [48] using best parameter combination using Jensen-Shannon (JS) divergence where T=1. The results were averaged over 5 runs in order to determine its stability. The results obtained were very impressive since they outperformed the R-GTCG approach except in 3 cases that is in the CoffeeBreak's precision **[Prec=0.5125 compare to ours Prec=0.49260]** and PosterData's precision and F1 **[Prec=0.8146 and F1=0.8354 compare to ours 0.78542 and 0.82296 respectively]**. The rest our new approach outperformed the R-GTCG approach.

We also carried out an evolution on the 15 sequences of the FiendsMeet2 dataset using the parameter as in [48] where $T = \frac{2}{3}$ (that is when two-third of the group members are detected) and T=1 (when all the group members are detected). Summary of the results are reported in Table 7.4 and Table 7.3. In the first case using $T = \frac{2}{3}$ Table 7.4, there is only a single scenario in Sequence 8 **[where Prec=0.68278,Rec=0.65559 and F1=0.66891 compare to ours Rec=0.67069,0.64048 and .065524 respectively]** in which the R-GTCG completely outperformed our new approach in all the 3 measures. This could be attributed to issues of the previous approach as discussed earlier. Despite R-GTCG outperforming our approach the difference is less than 2%, well below the variance found in the experiment. Other notable cases where our approach slightly trailed behind is in Sequence 10's precision **[Prec=0.97187 and F1=0.95097 compare to ours 0.95269 and 0.94495 respectively]** and F1 and Sequence 15 **[Prec=0.93551 compare to ours 0.93408** in which the variances are still below one found in the experiment. One notable improvement in our approach is in Sequences 11-13 which in the R-GTCG has very low performance. This

Tab. 7.2: Results on single frame experiment with the parameters as discussed in the previous chapters (σ in Eq.4.8 and l in Eq.4.7 using T=1)

CoffeeBreak(S1+S2)			
Method	**Prec**	**Rec**	**F1**
R-GTCG	**0.5125**	0.5163	0.5125
OURS	0.49260	**0.55090**	**0.52010**
l = *130*, $\sigma = 0.3$			

PosterData		
Prec	**Rec**	**F1**
0.8146	0.8573	**0.8354**
0.78542	**0.86427**	0.82296
l = *80*, $\sigma = 0.3$		

CocktailParty			
Method	**Prec**	**Rec**	**F1**
R-GTCG	0.3828	0.3828	0.3828
OURS	**0.4422**	**0.4490**	**0.4455**
l = *105*, $\sigma = 0.9$			

Synth		
Prec	**Rec**	**F1**
0.3000	0.2100	0.2471
0.7413	**0.7780**	**0.7592**
l = *70*, $\sigma = 0.1$		

Gdet			
Method	**Prec**	**Rec**	**F1**
R-GTCG	0.5126	0.5084	0.5105
OURS	**0.5297**	**0.5310**	**0.5304**
l = *145*, $\sigma = 0.7$			

Tab. 7.3: Results on the 15 Sequences of theFriendsMeet2 using T=1

	Seq1			Seq2			Seq3		
Method	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
R-GTCG	0.69527	0.67426	0.68460	0.90242	0.89107	0.89671	**0.74896**	**0.74896**	**0.74896**
Ours	**0.75482**	**0.75482**	**0.75482**	**0.92057**	**0.91528**	**0.91792**	**0.74896**	**0.74896**	**0.74896**

	Seq4			Seq5			Seq6		
Method	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
R-GTCG	0.61123	0.61019	0.61071	0.87119	0.84164	0.85616	0.88654	0.88654	0.88654
Ours	**0.81705**	**0.81705**	**0.81705**	**0.89381**	**0.89474**	**0.89427**	**0.89259**	**0.89259**	**0.89259**

	Seq7			Seq8			Seq9		
Method	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
R-GTCG	**0.80783**	0.79928	0.80353	0.18731	0.18731	0.18731	0.96184	0.95988	0.96086
Ours	0.80603	**0.80828**	**0.80715**	**0.89381**	**0.62538**	**0.62538**	**0.99804**	**0.99609**	**0.99706**

	Seq10			Seq11			Seq12		
Method	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
R-GTCG	0.88875	0.87596	0.88231	0.00948	0.00474	0.00632	**0.00000**	**0.00000**	**0.00000**
Ours	**0.90537**	**0.90537**	**0.90537**	**0.84360**	**0.84360**	**0.84360**	**0.00000**	**0.00000**	**0.00000**

	Seq13			Seq14			Seq15		
Method	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
R-GTCG	**0.03131**	**0.03131**	**0.03131**	0.08329	0.08329	0.08329	0.85840	0.83793	0.84804
Ours	0.00196	0.00196	0.00196	**0.08716**	**0.08716**	**0.08716**	**0.85878**	**0.88959**	**0.87392**

Tab. 7.4: Results on the 15 Sequences of theFriendsMeet2 using $T = \frac{2}{3}$

	Seq1			Seq2			Seq3		
Method	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
R-GTCG	0.88441	0.82137	0.85172	0.92537	0.93419	0.92976	**0.74896**	**0.74896**	**0.74896**
Ours	**0.94046**	**0.91506**	**0.92758**	**0.98336**	**0.97529**	**0.97931**	**0.74896**	**0.74896**	**0.74896**
	Seq4			**Seq5**			**Seq6**		
Method	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
R-GTCG	0.85863	0.7578	0.80507	0.90859	0.87165	0.88974	0.9289	0.9289	0.9289
Ours	**0.87526**	**0.85655**	**0.86580**	**0.91136**	**0.91274**	**0.91205**	**1.00000**	**1.00000**	**1.00000**
	Seq7			**Seq8**			**Seq9**		
Method	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
R-GTCG	0.91389	0.91854	0.91621	**0.68278**	**0.65559**	**0.66891**	**1.00000**	**0.99804**	**0.99902**
Ours	**0.97345**	**0.97885**	**0.97614**	0.67069	0.64048	0.65524	**1.00000**	**0.99804**	**0.99902**
	Seq10			**Seq11**			**Seq12**		
Method	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
R-GTCG	**0.97187**	0.93095	**0.95097**	0.41232	0.30806	0.35264	0.11231	0.13478	0.12252
Ours	0.95269	**0.93734**	0.94495	**0.8910**	**0.86493**	**0.87777**	**0.34942**	**0.65225**	**0.45506**
	Seq13			**Seq14**			**Seq15**		
Method	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**Prec**	**Rec**	**F1**
R-GTCG	0.12818	0.14090	0.13424	0.86818	0.86818	0.86818	**0.93551**	0.91408	0.92467
Ours	**0.27299**	**0.32681**	**0.29749**	**0.91043**	**0.91043**	**0.91043**	0.93408	**0.97761**	**0.95535**

is attributed to the fact that these sequences are not conversation groups but contains queues and people walking which produces a lot of false positives. Our approach performed extremely well especially in Sequence 11 with a mean value of **0.85** compared to the R-GTCG value of **0.40**. This is a significant improvement. In Sequences 12 and 13, we were able to double the performance. This is attributed to fact that, our appraoch was able to deal with senarios involving queues which was an issue with in the other approach as discussed earlier.

Furthermore, we did an evaluation on the same dataset using T=1 as reported in Table 7.3. Just like in the previous case where $T = \frac{2}{3}$, our approach still outperform R-GTCG except in one Sequence 13 **[Prec=0.03131,Rec=0.03131,F10.03131 compare with our 0.00196,0.00196 and 0.00196 respectively]** in which our method performed badly. The rest it performed extremely well even in case where it is slightly trailed considering the variance which is very low.

8. CONCLUSION

In this work, we proposed a new approach which addresses the problems of of queues, by checking if two persons are facing or not using cousine similarity. We introduce this to act as a filter where frustum overlaps detects scenarios where persons are queueing as conversational groups since queues produces a lot of false positives. Furthermore, we also found a solution to deal with the high positive values generated by the use of Jensen-Shannon (JS) divergence to probability distributions that do not have anything in common. Our new approach improves upon the existing Robust Game -Theory for Conversational Groups which captures the pairwise scores between individuals by stochastically modelling the social attention of a group. The affinity matrix is filled by the pairwise scores which encodes an edge weighted graph. This encoded matrix represents the whole scenery under preview. Furthermore, in order to finds groups in the scene, a game-theoretical clustering strategy is used which defines the objects to be grouped as strategies corresponds to the detected individual and the payoff function corresponds to similarity measure. This strategy has been proven to be efficient when compared to other strategies. Our new approach results show significant improvements when compared to Robust Game -Theory for Conversational Groups on the single-frames situation on six publicly available datasets.

BIBLIOGRAPHY

[1] *Vision-based pedestrian detection: the PROTECTOR system*, 2004.

[2] *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*. IEEE Computer Society, 2006.

[3] Sebastiano Battiato and José Braz, editors. *VISAPP 2014 - Proceedings of the 9th International Conference on Computer Vision Theory and Applications, Volume 2, Lisbon, Portugal, 5-8 January, 2014*. SciTePress, 2014.

[4] L. Bazzani, D. Tosato, M. Cristani, M. Farenzena, G. Pagetti, G. Menegaz, and V. Murino. Social interactions by visual focus of attention in a three-dimensional environment. *Expert Systems*, 2013.

[5] Samuel R. Bulò and Marcello Pelillo. A game-theoretic approach to hypergraph clustering. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1571–1579. Curran Associates, Inc., 2009.

[6] Cheng Chen and Jean-Marc Odobez. We are not contortionists: Coupled adaptive learning for head and body orientation estimation in surveillance video. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2012.

[7] Wongun Choi, Yu-Wei Chao, Caroline Pantofaru, and Silvio Savarese. Discovering groups of people in images. In *European Conference on Computer Vision*, pages 417–433. Springer International Publishing, 2014.

[8] T. Matthew Ciolek and Adam Kendon. Environment and the spatial arrangement of conversational encounters. *Sociological Inquiry*, 50(3-4):237–271, 1980.

[9] R. T. Collins, Weina Ge, and R. B. Ruback. Vision-based analysis of small groups in pedestrian crowds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(5):1003–1016, 2012.

[10] M Cristani, G Paggetti, A Vinciarelli, L Bazzani, G Menegaz, and V Murino. Towards computational proxemics: Inferring social relations from interpersonal distances. In *Proceedings of IEEE International Conference on Social Computing*, pages 290–297, 2011.

[11] Marco Cristani, Loris Bazzani, Giulia Paggetti, Andrea Fossati, Diego Tosato, Alessio Del Bue, Gloria Menegaz, and Vittorio Murino. Social interaction discovery by statistical analysis of f-formations. In Hoey et al. [22], pages 1–12.

[12] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 1 - Volume 01*, CVPR '05, pages 886–893, Washington, DC, USA, 2005. IEEE Computer Society.

[13] Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, September 2010.

[14] D.R. Forsyth. *Group dynamics*. Thomson/Wadsworth, 2006.

[15] Tian Gan, Yongkang Wong, Daqing Zhang, and Mohan S. Kankanhalli. Temporal encoded f-formation system for social interaction detection. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 937–946, New York, NY, USA, 2013. ACM.

[16] H Garfinkel. *Studies in Ethnomethodology*. Prentice-Hall, Englewood Cliffs, New Jersey, 1967.

[17] E. Goffman. *Encounters: two studies in the sociology of interaction*. The advanced studies in sociology series. Bobbs-Merrill, 1961.

[18] Erving Goffman. *Behavior in Public Places: Notes on the Social Organization of Gatherings*. Free Press, September 1966.

[19] G. Groh, A. Lehmann, J. Reimers, M. R. Friess, and L. Schwarz. Detecting social situations from interaction geometry. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pages 1–8, Aug 2010.

[20] E. T. Hall. *The Hidden Dimension: Man's Use of Space in Public and Private.* The Bodley Head Ltd, 1966.

[21] E.T. Hall. *The hidden dimension*. Doubleday Anchor Books. Doubleday, 1966.

[22] Jesse Hoey, Stephen J. McKenna, and Emanuele Trucco, editors. *British Machine Vision Conference, BMVC 2011, Dundee, UK, August 29 - September 2, 2011. Proceedings*. BMVA Press, 2011.

[23] Hayley Hung and Ben Kröse. Detecting f-formations as dominant sets. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, ICMI '11, pages 231–238, New York, NY, USA, 2011. ACM.

[24] Varun Jain and James L. Crowley. Head Pose Estimation Using Multi-scale Gaussian Derivatives. In *18th Scandinavian Conference on Image Analysis*, Espoo, Finland, June 2013.

[25] Adam Kendon. *Conducting interaction: Patterns of behavior in focused encounters*. Cambridge University Press, 1990.

[26] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori. Discriminative latent models for recognizing contextual group activities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(8):1549–1562, Aug 2012.

[27] Ruonan Li, Parker Porfilio, and Todd Zickler. Finding group interactions in social clutter. 2013.

[28] J. Lin. Divergence measures based on the shannon entropy. *IEEE Trans. Inf. Theor.*, 37(1):145–151, September 2006.

[29] M. J. Marin-Jimenez, A. Zisserman, M. Eichner, and V. Ferrari. Detecting people looking at each other in videos. *International Journal of Computer Vision*, 106(3):282–296, 2014.

[30] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *European Conference on Computer Vision*. Springer-Verlag, 2004.

[31] Anuj Mohan, Constantine Papageorgiou, and Tomaso Poggio. Example-based object detection in images by components. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:349–361, 2001.

[32] Rensso V. H. Mora-Colque, Guillermo Cámara-Chávez, and William Robson Schwartz. *Detection of Groups of People in Surveillance Videos Based on Spatio-Temporal Clues*, pages 948–955. Springer International Publishing, Cham, 2014.

[33] T. S. Motzkin and E. G. Straus. Maxima for graphs and a new proof of a theorem of Turán. *Canadian Journal of Mathematics*, 17:533–540, 1965.

[34] Erik Murphy-Chutorian and Mohan Manubhai Trivedi. Head pose estimation in computer vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(4):607–626, April 2009.

[35] John Nash. Non-cooperative games. *Annals of Mathematics*, 54(2):286–295, 1951.

[36] Constantine Papageorgiou and Tomaso Poggio. A trainable system for object detection. *Int. J. Comput. Vision*, 38(1):15–33, June 2000.

[37] M. Pavan and M. Pelillo. Dominant sets and hierarchical clustering. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 362–369 vol.1, Oct 2003.

[38] Massimiliano Pavan and Marcello Pelillo. Dominant sets and pairwise clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(1):167–172, January 2007.

[39] Sudeep Sarkar and Kim L. Boyer. Quantitative Measures of Change based on Feature Organization: Eigenvalues and Eigenvectors. In *CVPR '96: Proceedings of the 1996 Conference on Computer Vision and Pattern Recognition (CVPR '96)*, Washington, DC, USA, 1996. IEEE Computer Society.

[40] Henry Schneiderman and Takeo Kanade. Object detection using the statistics of parts. *Int. J. Comput. Vision*, 56(3):151–177, February 2004.

[41] David Schweingruber and Clark McPhail. A method for systematically observing and recording collective action. *Sociological methods & research*, 27(4):451–498, 1999.

[42] F. Setti, O. Lanz, R. Ferrario, V. Murino, and M. Cristani. Multi-scale f-formation discovery for group detection. In *2013 IEEE International Conference on Image Processing*, pages 3547–3551, Sept 2013.

[43] Francesco Setti, Chris Russell, Chiara Bassetti, and Marco Cristani. F-formation detection: Individuating free-standing conversational groups in images. *CoRR*, abs/1409.2702, 2014.

[44] Andrea Torsello, Samuel Rota Bulò, and Marcello Pelillo. Grouping with asymmetric affinities: A game-theoretic perspective. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA* [2], pages 292–299.

[45] Khai N. Tran, Apurva Bedagkar-gala, Ioannis A. Kadadiaris, and Shishir K. Shah. Social cues in group formation and local interactions for collective activity analysis.

[46] Khai N. Tran, Apurva Gala, Ioannis A. Kakadiaris, and Shishir K. Shah. Activity analysis in crowded environments using social cues for group discovery and human interaction modeling. *Pattern Recognition Letters*, 44:49–57, 2014.

[47] G. Tzanetakis and N. O'Connor, editors. *Group Detection in Still Images by F-formation Modeling: A Comparative Study*, Paris, 2013. IEEE.

[48] Sebastiano Vascon, Eyasu Zemene Mequanint, Marco Cristani, Hayley Hung, Marcello Pelillo, and Vittorio Murino. Detecting conversational groups in images and sequences: A robust game-theoretic approach. *Computer Vision and Image Understanding*, 143:11–24, 2016.

[49] Paul Viola, Michael J. Jones, and Daniel Snow. Detecting pedestrians using patterns of motion and appearance. *Int. J. Comput. Vision*, 63(2):153–161, July 2005.

[50] T. Yu, S. N. Lim, K. Patwardhan, and N. Krahnstoever. Monitoring, recognizing and discovering social networks. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1462–1469, June 2009.

[51] Eyasu Zemene, Samuel Rota Bulò, and Marcello Pelillo. *Dominant-Set Clustering Using Multiple Affinity Matrices*, pages 186–198. Springer International Publishing, Cham, 2015.

[52] Gloria Zen, Bruno Lepri, Elisa Ricci, and Oswald Lanz. Space speaks: Towards socially and personality aware visual surveillance. In *Proceedings of the 1st ACM International Workshop on Multimodal Pervasive Video Analysis*, MPVA '10, pages 37–42, New York, NY, USA, 2010. ACM.