Predicting Group Convergence in Egocentric Videos

Jyoti Nigam and Renu M. Rameshan
Indian Institute of Technology, Mandi, Himachal Pradesh, India

Keywords: Ego-vision, Group Convergence, Gaze Mask, Strongly Connected Graph.

Abstract: In this work, our aim is to find the group dynamics in social gathering videos that are captured from the first-person perspective. The complexity of the task is high as only one sensor (wearable camera) is present to sense all the N agents. An additional complexity arises as the first-person who is part of the group is not visible in the video. In particular, we are interested in identifying the group (with certain number of agents) convergence. We generate a dataset named *EgoConvergingGroup* to evaluate our proposed method. The proposed method predicts the group convergence in 90 to 250 number of frames, which is much ahead of the actual convergence.

1 INTRODUCTION AND RELATED WORK

Over the past few years wearable cameras have become very common and analysis of egocentric videos has become a well recognized research area. We focus on egocentric videos involving social interactions. Our aim is to analyze and predict group formation in such videos. Activities like people converging to form a group, diverging from a group, standing together, *etc* (Bhargava and Chaudhuri, 2014) are classified as social interactions. Recognizing such group interactions in a social gathering where first-person is also a part of that group is a challenging and interesting problem which also has various applications.

A prime application is in mining the interesting moments and group activities from the videos captured in the entire day. The information about the different types of interactions in a social gathering can aid in various other applications like: (i) finding connections among the agents, (ii) identifying who is meeting whom and recognition of isolated person, and (iii) finding center of attraction in the group along with interacting individuals. The above mentioned applications have a significant role in surveillance, safety and social behavior analysis (Bhargava and Chaudhuri, 2014).

In (Alletto et al., 2015), the authors detected stationary (already formed) groups in ego-vision scenarios. People in the scene are tracked through out the video and their head pose and 3D locations are estimated. In (Bhargava and Chaudhuri, 2014), the group dynamics is captured by a single surveillance camera which is fixed at a certain height and sees all the agents throughout which can give complete trajectory information of the group.

In egocentric videos due to the absence of a third-person static camera we cannot get the global information *i.e.* all agents are not being captured from the camera as the first-person is not visible in the videos. In the proposed work, we consider a social gathering scenario in which the agents are moving towards each other to form a group. We are particularly interested in finding the criterion for group convergence.

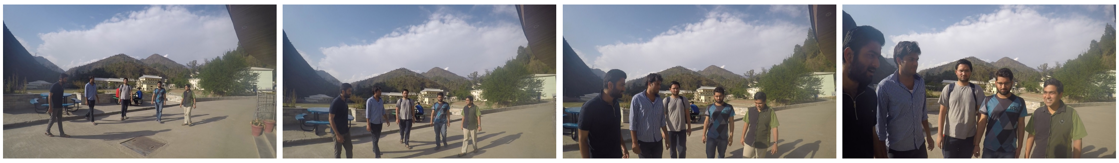
To handle the above mentioned challenges, we adopted the technique from (Lin et al., 2004), which addresses the problem of predicting group convergence of mobile autonomous agents. Their solution is based on local distributed control strategy where each agent has a sensor which senses the neighboring agents to form the interaction graph. These graphs are analyzed to predict the group convergence.

We solve the problem of predicting group convergence in egocentric videos where we have only one sensor that is the first-person camera. We extract all the information needed for creating the graph from the video assuming all agents are visible throughout the video. The single sensor (camera) makes this problem challenging.

The contributions of the proposed work are

- Group dynamics of agents in egocentric videos are analyzed with only partial information about the group. The meaning of partial information is that out of N agents only $N-1$ agents are seen throughout the video.
- We show that the gaze information of an agent can

Nigam, J. and Rameshan, R.
Predicting Group Convergence in Egocentric Videos.
DOI: 10.5220/0007556207730777
In *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2019)*, pages 773-777
ISBN: 978-989-758-351-3



(a) frame 10 (b) frame 120 (c) frame 250 (d) frame 380

Figure 1: A group formation.

be used to establish its connection with others.

- We estimate via experimentations, the required time interval in which the interaction graphs should be strongly connected.

2 PROPOSED METHOD

Our analysis is restricted to the scenario where there is a social gathering in a region with one of the agents having wearable camera mounted on the head. We assume that the people/agents are initially scattered and eventually all of them move to form a group. Our goal is to predict whether the group including the first-person is converging to a common location or not. The prediction is done after observing few of the initial frames; let F be the set of these frames. In our experiments the size of F varies from 90 to 250 frames.

We generate a directed graph at every frame with agents as vertices and the relative gaze masks of all agents are used to estimate the connection among them. Once we have a directed graph, we check if the graph is strongly connected or not. For this we compute the transition matrices and their product over a time interval. These matrix products calculated for different intervals are analyzed (up to a certain number of frames) to predict the convergence of the group.

2.1 Tracking of Agents

We need to track all agents across the frames in F to obtain their gaze masks. To detect the faces, *Tiny Face* (Hu and Ramanan, 2017) method is used which can handle scale and view angle variation within a frame. All the faces, present in the first frame, are detected and provided as ground truth to the tracker.

The multi-target tracker tracks them in subsequent frames. We use the Real-time, Recurrent, Regression-based tracker, or *Re3* (Gordon et al., 2017) which is a fast and accurate network for generic multi-object tracking to do this task. To prevent the tracker from deviating while tracking the faces, *Tiny Face* module is applied at every *tenth* frame to re-initialize the tracker.

2.2 Gaze Mask Estimation

In a group of agents, the gaze mask of an individual implicitly encodes the connection with the surrounding agents. Hence we obtain the gaze mask for each agent. The gaze mask can be seen as the field of view of an agent and other agents who come in the range of that mask are considered as connected to that agent under consideration.

We use the network in (Recasens et al., 2017) with few modifications to obtain the gaze information of each tracked face. Instead of using face recognition method provided in the network we give the tracked faces along with their locations as input at each frame.

This network has two different pathways, one for saliency and the other for gaze. At each frame we provide input in the form of pairs of the full image and the cropped faces of each agent (output of *Re3* tracker) sequentially. The saliency pathway looks at the full image independently but does not know the agent's location, and gives a spatial map, $S(I_t)$ of size $W \times W$, where I_t is the current input frame.

The gaze pathway has access to only the closeup image of the agent's head and its location, that gives another spatial map, $g_t^i(a_i^h, a_i^l)$ where a_i^h and a_i^l are i^{th} agent's head image and location, respectively. This spatial map is also of size $W \times W$ (same as saliency pathway). The outputs of the two pathways are then integrated by an element-wise product to obtain gaze mask of each agent, as shown in Eq.(1).

$$\mathcal{G}_t^i = S(I_t) \times g_t^i(a_i^h, a_i^l), \tag{1}$$

where $\mathcal{G}_t^i$ is the gaze mask of i^{th} agent at t^{th} frame.

2.3 Interaction Graph and Neighborhood

In this step, we use a strategy for group convergence that is motivated by Reynolds' cohesion steering strategy (Lin et al., 2004). There are $N-1$ agents who are visible in the first-person camera and numbered 1 through $N-1$ and the first-person is the N^{th} agent. The agents are labeled and the tracking method takes care of their identities.

Each agent has a cone-like field of view (gaze mask) and those agents who overlap with the gaze mask of the i^{th} agent, are considered as neighbors of the i^{th} agent. Since it is not necessary that any two neighboring agents are looking at each other at the same time, we use a directed graph to model the connection among agents.

Once we have the gaze mask $\mathcal{G}_t^i$ of each agent a_t^i, $i \in (1,N)$, the neighborhood of an agent (represented by an N tuple), is populated as follows. Let $O(a_t^i)$ be the overlap between image I_t and gaze mask of i^{th} agent $\mathcal{G}_t^i$ as shown in Eq. (2),

$$O(a_t^i) = I_t \cap \mathcal{G}_t^i. \tag{2}$$

If $a_t^i \in O(a_t^i)$ then $\mathcal{N}(a_t^i) = 1$. Here $\mathcal{N}(a_t^i)$ is the neighborhood of i^{th} agent.

This process is repeated for each agent and an $N \times N$ connection matrix is formed with entries as $0/1$. All diagonal elements are zero assuming an agent cannot look at itself. In Fig. 2, we show an example of an interaction graph, obtained for one of the frames in our experimentation.

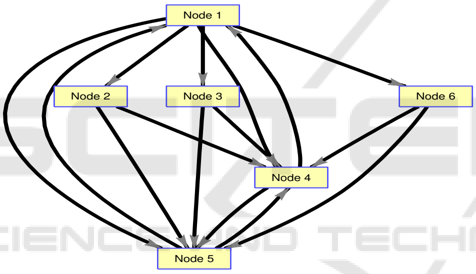


Figure 2: Connection graph representation based on agents' gaze mask, node1 is first-person who is looking at each agent.

2.4 Entries in the Connection Matrix

Once the interaction graph is formed we have an $N \times N$ connection matrix. Row and column indices of this matrix indicate agents. In a connection matrix a particular row indicates an agent's field of view (out of N agents how many are visible to it) and column values reflect the identities of agents who are looking at that particular agent.

Two cases need special consideration. First is the neighbors of first-person, in this case the neighborhood vector is all *ones*. Since all agents are seen by first-person according to our assumption. The second one is that in which an agent is looking at the first-person (camera). Since, in the gaze estimation procedure the gaze cone is originated at the middle point between the eyes, the gaze direction comes orthogonal to the agent's face. We identify this case by checking the salient objects coming in the overlap of the gaze mask of that agent and if the face or a significant portion (approximately half) of the face comes in the overlap, we infer that the agent is looking at the first-person, an example is shown in Fig. 3.

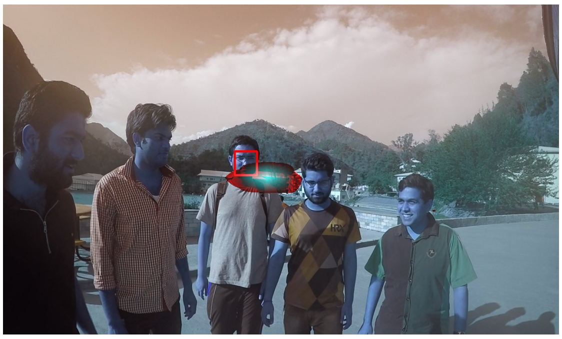


Figure 3: Gaze mask of an agent who is looking at first-person.

2.5 Strong Connectedness

From the above sections we can see that a directed graph is generated at each frame. Since, there are N agents present in the group they can be connected in $\mathcal{P} = 2^{n^2-n}$ possible ways. Let $\{G_p : p \in \mathcal{P}\}$ be the set of directed graphs. For every directed graph G_p, C_p denotes the adjacency/connection matrix and D_p denotes the diagonal matrix where each i^{th} diagonal element shows the number of all outgoing directed edges from node i. R_p denotes the diagonal matrix whose i^{th} diagonal element is the reciprocal of the i^{th} diagonal element of matrix D_p if it is not zero, and zero if it is zero. Then, we define $A_p = R_p(C_p - D_p)$.

A transition matrix is defined as $\Phi(t,t_i) = e^{A_{p(t_i)}(t-t_i)}$ for a time interval $t - t_i$ where $t \in [t_i, t_{i+1}]$. This is a row stochastic matrix (Lin et al., 2004). The product of transition matrices from set $\{\Phi(t,t_i)\Phi(t_i,t_{i-1})\ldots\Phi(t_1,t_0)\}$ is denoted by Ψ_t at time t. The group convergence is predicted from a sequence of such row stochastic matrices according to Theorem 1. The steps for convergence analysis are given in Algo. 1. The condition for convergence according to Theorem 4 (Lin et al., 2004), is that the connection matrix should be connected in a time interval of T. The condition for convergence reduces to $\lambda(\Psi) < 1$ and eigenvalue of Ψ should be 1 with algebraic multiplicity *one* for single group.

We reproduce the theorem mentioned in (Lin et al., 2004) for convenience of the reader.

Theorem 1. *Let $\{M_1, M_2 \ldots\}$ be a finite or infinite set of row stochastic matrices satisfying $0 \leq \lambda(M_i) \leq \beta < 1$. Then for each infinite sequence, $M_{k_1}, M_{k_2}, \ldots$, there exists a row vector r such that*

Table 1: Moving pattern shows how agents are moving and looking (including first-person). T shows the time period in which the graph gets strongly connected. Algebraic multiplicity is denoted as AM.

Case	Moving Pattern	Number of videos	T	$\lambda(\Psi)$	Top eigenvalue	Convergence?
1	Seven agents are moving towards the center of the group and looking at each other frequently and first-person is looking at each agent throughout	5	connected throughout	0.8 to 0.9	1 (AM=1)	Yes (4/5). Prediction at 60^{th} frame and actual convergence at 150^{th} frame.
2	Ten agents are moving towards the center of the group and looking at each other occasionally and first-person is looking at each agent throughout	10	2-5 (sec)	0.7 to 0.9	1 (AM=1)	Yes (8/10). Prediction at 190^{th} frame and actual convergence at 250^{th} frame.
3	Fifteen agents are moving randomly and looking at each other occasionally and first-person is looking at limited number of agents at once	10	Not connected	N/A	1 ($AM > 1$)	NO

$\lim_{j\to\infty} M_{k_j}, M_{k_{j-1}} \ldots M_{k_1} = \mathbf{1}r$, *where* $\mathbf{1}$ *is a* N *dimensional vector of ones.*

Here, $\lambda(M)$ is defined as:

$$\lambda(M) = 1 - \min_{i_1, i_2, i_1 \neq i_2} \sum_j \min(m_{i_1 j}, m_{i_2 j}). \tag{3}$$

Algorithm 1: Group Convergence.

```
Require:
    Sequence of $(N \times N)$ transition matrices $\Phi(t)$ for each frame in set $F$
Ensure:
    Prediction of convergence
for $t = 1 : T : \text{size}(F)$ do
    for $i = 1 : T$ do
        $\Psi_i = \text{Product}(\Phi(i), ..., \Phi(1))$
    end for
    Compute $\lambda(\Psi_t)$
    if $\lambda(\Psi_t) < 1$ then
        group converges
    else
        group does not converge
    end if
end for
```

3 EXPERIMENTATION AND RESULTS

Given an egocentric video, our first task is to find the number of agents in the group and it is done by using Tiny Face method in the first frame (assuming all agents are visible in the first frame). These detected faces are labeled from 1 to $N-1$ and fed as input to Re3 tracker along with their identities. The tracker predicts the location of each face in the subsequent frames (belong to the set F). These faces and their locations are processed to estimate their gaze masks. The $N \times N$ connection matrix (treating field of view of the camera as the gaze mask of first-person) at each frame is formed by using the neighborhood of each agent which is obtained by gaze mask. These connection matrices are used to find transition matrices over a time interval (5 sec in the proposed experimentations). The product of these transition matrices is denoted as Ψ.

We analyze the eigenvalues and λ (as defined in Eq. (3)) of this matrix product Ψ in each time interval, the range of these values are given in Tab. 1. The largest eigenvalue with algebraic multiplicity (AM) *one*, shows the presence of a single group where greater AM reflects the presence of more than one groups. Since we are considering a single group in our experimentation, we expect eigenvalue 1 with $AM = 1$. The values of $\lambda(\Psi)$ remain less than *one* thereby satisfying the criterion mentioned in Theorem 1.

We give the range of the time interval (T) in which the digraph becomes strongly connected via experiments. In the first category of videos, the graphs remain strongly connected throughout. The group convergence is predicted until any agent moves out of the frame. In the next category, we observed the time interval $2-5$ *sec*. We keep predicting the convergence until the graphs remain strongly connected throughout. In the last category, the graphs never became strongly connected.

Dataset. We choreographed approximately twenty videos, each of $1-2$ minutes duration. To validate the proposed approach we generate three types of video. In all the videos agents are scattered in the initial frames. In first category the agents are looking at each other frequently and walking towards a common point to form a single group. In the next category the agents are looking at each other after a significant time period and forming a single group. In the last category the agents are not looking at each other and remain scattered till the end of video. The details of the experimentation are given in Tab. 1.

There are some existing datasets (EGO-GROUP and EGO-HPE) (Alletto et al., 2014) which contain group of people in ego-vision but they are already formed and stationary, hence we do not consider these datasets for our experimentation.

We could not provide a comparison of our proposed method with existing methods as we are focusing on formation of a group and not on the already formed groups. In existing works considering a single or multiple groups, the membership of agents are obtained but in this work we consider only a single group with fixed number of agents.

4 CONCLUSION

In this paper, we proposed a method to find the convergence of a group containing a single first-person moving camera. We show via experimentations that global information and communication are not required for solving group dynamics problem. Instead local estimates *i.e.* the gaze masks of each agent extracted from the recorded video, used to predict global group convergence.

ACKNOWLEDGEMENT

This publication is an outcome of the R & D work undertaken project under the Visvesvaraya PhD Scheme of Ministry of Electronics and Information Technology, Government of India, being implemented by Digital India Corporation

REFERENCES

Alletto, S., Serra, G., Calderara, S., and Cucchiara, R. (2015). Understanding social relationships in egocentric vision. *Pattern Recognition*, 48(12):4082–4096.

Alletto, S., Serra, G., Calderara, S., Solera, F., and Cucchiara, R. (2014). From ego to nos-vision: Detecting social relationships in first-person views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 580–585.

Bhargava, N. and Chaudhuri, S. (2014). Finding group interactions in social gathering videos. In *Proceedings of the 2014 Indian Conference on Computer Vision Graphics and Image Processing*, page 5. ACM.

Gordon, D., Farhadi, A., and Fox, D. (2017). Re3: real-time recurrent regression networks for object tracking. *arXiv preprint arXiv:1705.06368*, 2.

Hu, P. and Ramanan, D. (2017). Finding tiny faces. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 1522–1530. IEEE.

Lin, Z., Broucke, M., and Francis, B. (2004). Local control strategies for groups of mobile autonomous agents. *IEEE Transactions on automatic control*, 49(4):622–629.

Recasens, A., Vondrick, C., Khosla, A., and Torralba, A. (2017). Following gaze in video. In *The IEEE International Conference on Computer Vision (ICCV)*, volume 4.