INTERACTION-GCN: A GRAPH CONVOLUTIONAL NETWORK BASED FRAMEWORK FOR SOCIAL INTERACTION RECOGNITION IN EGOCENTRIC VIDEOS

Simone Felicioni

University of Perugia
simone.felicioni@studenti.unipg.it

Mariella Dimiccoli

Institut de Robòtica i Informàtica Industrial (CSIC-UPC)
mdimiccoli@iri.upc.edu

ABSTRACT

In this paper we propose a new framework to categorize social interactions in egocentric videos, we named InteractionGCN. Our method extracts patterns of relational and nonrelational cues at the frame level and uses them to build a relational graph from which the interactional context at the frame level is estimated via a Graph Convolutional Network based approach. Then it propagates this context over time, together with first-person motion information, through a Gated Recurrent Unit architecture. Ablation studies and experimental evaluation on two publicly available datasets validate the proposed approach and establish state of the art results.

Index Terms— social interaction recognition, graph convolutional networks, egocentric vision

1. INTRODUCTION

Recently, wearable cameras have enabled the automatic capture of social life in a naturalistic setting, from a first-person point of view [1]. This has opened the unique opportunity of analyzing the real involvement in social interactions at the personal level [2–4]. However, the research focus in the egocentric vision domain has been so far on the detection [3, 4] and classification of social interactions based on the kind of relations [5, 6], while the problem of fine-grained classification of a specific relation based on the degree of interactivity has been addressed only in [2]. This latter categorization would be of paramount importance to truly understand social interactions and thus to allow an easier human-machine communication. However, it faces several challenges. Firstly, compared to social relation classification, a categorisation based on the degree of interactivity is more fine-grained and ambiguous. In particular, a model must be able to discriminate whether it is an interactive exchange (discussion or dialogue depending on many people or just two are actively involved, even in presence of multiple persons around) or largely one-sided with a single person speaking most of the time (monologue). Secondly, since the camera is worn in a naturalistic setting, interaction cues present a large intra-class variability as well as unpredictable camera/head movement. Fig. 1 illustrates the difficulty of the problem in the GeorgiaTech Social Interaction dataset[1] [2], captured by a head-mounted wearable camera in an amusement park: five different conversation types may occur at different time intervals during a social interaction, even in absence of drastic visual changes, e.g, discussion vs dialogue. In [2], this problem was addressed by using a Hidden Conditional Random Field (HCRF) formulation that accounts for the dependencies between state labels over time. However, it fails in modelling the interactivity context at the frame level.

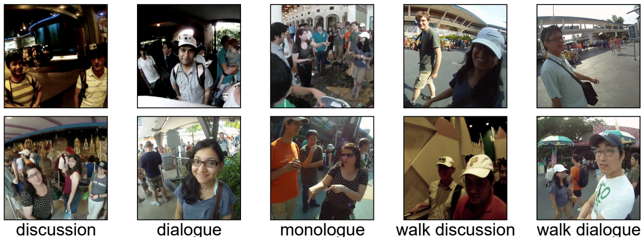


Fig. 1. Different categories in the GeorgiaTech dataset [2]
.

In this paper, we overcome this limitation by explicitly modelling the interactivity context by building a relational graph, where nodes correspond to persons with their associated individual features, and edges correspond to interaction cues between pairs of persons. We learn the interactivity context from each frame as embedding on this relational graph via a Relational-Graph convolutional network (R-GCN). It is then fed to a Gated Recurrent Unit (GRU) [7] together with first-person motion. A visual overview of our InteractionGCN framework is given in Fig. 2. Experimental results on two publicly available datasets [2, 5] validate the proposed approach.

2. RELATED WORK

Third-person social interaction recognition A number of works have undertaken the problem of recognizing social relations on still images [8, 9], or particular family relations

Work partially funded by projects MINECO/ERDF RyC, PID2019-110977GA-I00, RED2018-102511-T, and MdM-IP-2019-03.

[1] http://cbs.ic.gatech.edu/egocentric/datasets.htm

in personal photo albums [10–13]. More recently, a formalization of people social life based on the Bugental theory has been proposed [14]. The literature in the video domain is much more scarce. A body of work has focused on the detection of *social roles* of people when interacting in videos [15–17] by relying on relative age, gender, clothing and again on people relative location. Recently, [18] proposed a Multiscale Spatial-Temporal Reasoning framework to classify social interactions from videos into the eight more common subcategories among the ones proposed in [14].

First-person social interaction recognition So far most efforts have focused on detecting groups of interacting people [19], detecting social saliency [20, 21], or detecting with whom the camera wearer is interacting [4, 22]. Only a few works have gone beyond the detection task in the egocentric domain [5, 6, 23]. Aghaei et al. [6] introduced a pipeline for automatic analysis of duration, frequency, type of relation, and diversity of the social interactions of a user captured by a wearable photo-camera during several weeks. A detailed classification of social relations in the egocentric domain, based on the Bugental's theory, has been recently proposed in [5]. Similarly to us, [2] aims at the classification of social interactions into five types depending on the modality of interaction. This is achieved through a HCRF based formulation. Although this model has shown encouraging results, it fails in capturing the conversational context at the frame level.

Graph convolutional networks Recently, Graph Convolutional Networks (GCN) have shown promising results in a variety of problems requiring the manipulation of graph-structured data, including several computer vision tasks [24, 25]. The key idea underlying GCN is the propagation of node information through message passing among neighbor nodes that is achieved by aggregating node information. Although the success of graph-based modelling for social interaction understanding [2, 9, 10, 12, 15, 16], GCN have been little explored in this context so far [18]. However, this work focus on recognizing the type of social relation, e.g, colleagues vs father-child for which objects, and person appearance play an important role. In contrast, we aim at a more fine-grained classification in terms of dialogue, discussion, monologue, etc. as originally proposed in [2], independently on the kind of relation.

3. PROPOSED APPROACH

We propose a new approach to classify social interactions in egocentric videos, which follows the taxonomy introduced in [2]. The overall architecture of our InteractionGCN framework is shown in Figure 2, and it is detailed below.

3.1. Feature extraction

For each of M persons in the scene, say $[P_1, ..., P_M]$, two different types of features are extracted: non-relational and relational features. The former take into account information that belongs to a specific person, independently of other interacting people. The latter take into account interaction features between pairs of persons. Since the camera is head-mounted, the camera wearer is assumed to look at the center of the scene and to be located at distance zero from the camera.

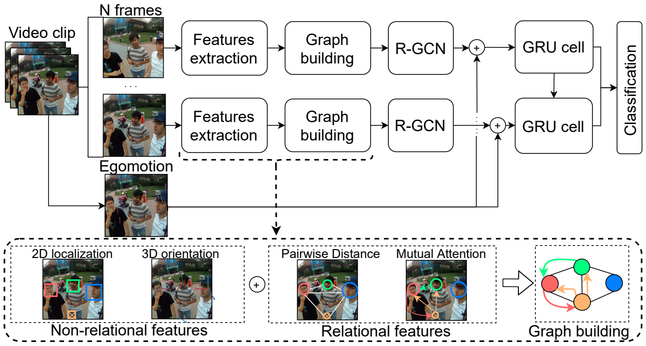


Fig. 2. Overview of the InteractionGCN framework.

Non-relational features To estimate non-relational features (i.e. 3D head orientation, location, first-person motion), in each selected frame from a video clip we detected faces, grouped them across frames and estimated their 3D orientation and 2D localization by using the Microsoft Azure Cognitive Systems API [2]. Finally, first-person motion features were estimated by using the method proposed in [26], that computes homography matrices for each pair of consecutive frames.

Relational features To estimate directional attention, we assumed that a person's gaze is strongly correlated to his/her 3D head orientation and can be described as a cone in 3D space, so that a person P_i is looking at a person P_j if the cone spanned by the head orientation of P_i includes P_j. The relative distance between P_i and P_j was computed by relying on the pre-estimated 2D distance of every person from the camera in a bird-view model. Similarly to [2], the latter was obtained by fitting a polynomial regression model built previously by collecting values of the height of faces in the image at different distance to a tripod mounted GoPro camera.

3.2. InteractionGCN framework

We generate a graph $\mathcal{G} = (V, E, R, \mathcal{W})$ for each frame, where V is the set of nodes, E the set of edges, R the set of relations represented by the edges (i.e. distance and attention), $\mathcal{W}$ the set of weights. Each node $v_i \in V$ corresponds to a person P_i having as node features non-relational cues f_i, e.g. stacked 3D head orientation and 2D localization vectors. Graph nodes are connected by two type of edges ($e^r_{ij} \in E, r \in R = \{d, a\}$), corresponding to two different pairwise relations: relative distance between pairs of individuals

[2] https://azure.microsoft.com/en-us/services/cognitive-services/

in the scene (d), including the camera wearer, and directional attention pattern (a). Distance edges (e^d_{ij}) are bi-directional and represent the distance between P_i and P_j, while attention edges (e^a_{ij}) are directed edges pointing from P_i to P_j, meaning that P_i is looking at P_j. $\alpha^r_{ij} \in \mathcal{W}$ is the weight of the labeled edge e^r_{ij}. We feed this graph initialized with the computed relational and non-relational features to a R-GCN [27] that propagates information among nodes in a local graph neighborhoods, yielding a richer representation relevant for interaction recognition, as it will be shown in Section 4. The propagation model used to update a node v_i in a R-GCN is an accumulated transformed feature vectors of neighboring nodes obtained as follows:

$$h_i^{(k+1)} = \sigma\left(\sum_{r\in\mathcal{R}}\sum_{j\in\mathcal{N}_i^r} \frac{\alpha^r_{ij}}{c_{i,r}} W_r^{(k)} h_j^{(k)} + \alpha^r_{ii} W_0^{(k)} h_i^{(k)}\right)$$

where $h^{(k+1)}$ is the hidden state of node v_i at the $k+1$ layer in a R-GCN, σ is the ReLU activation function, W_r and W_0 are learnable parameters of the transformation, α_{ij} and α_{ii} are the edge weights, $\mathcal{N}_i^r$ denotes the set of neighbor indices of node i connected by a relation type $r \in \mathcal{R}$, while $1/c_{i,r}$ is a normalization constant, ensuring that the node i receives a total weight contribution of 1, to which we assigned the value $c_{i,r} = |\mathcal{N}_i^r|$. This model accumulates transformed features of neighboring nodes through a normalized sum. More specifically, the node features are subject to relation-specific transformations, depending on the type and direction of an edge. Single self-connection of a special relation type to each node in the data ensures that the representation of a node at layer $l+1$ can also be informed by the corresponding representation at layer l. The output of the GCN at frame t is a list of node representations g_t. This list is vectorized and then concatenated with dynamic features corresponding to first-person motion estimation at frame t in a vector x_t. The collection of vectors for each frame is the input of a GRU [7], which captures the context information during all observations. The hidden state H is updated every frame t as follows: $H_{t+1} = GRU(H_t, x_t)$. More precisely:

$$z_t = \sigma(W_{xz}\cdot x_t + W_{Hz}\cdot H_t + b_z)$$

$$r_t = \sigma(W_{xr}\cdot x_t + W_{Hr}\cdot H_t + b_r)$$

$$\tilde{H} = tanh(W_{xH}\cdot x_t + W_{HH}\cdot(r_t * H_t) + b_H)$$

$$H_{t+1} = z_t * H_t + (1 - z_t) * \tilde{H},$$

where x_t and H_{t+1} are the input and output vectors, z_t and r_t are the update and reset vectors, W_s and b_s are the parameter matrices and bias vectors. The gate vector z_t is the trade-off parameter for updating hidden state H_{t+1} from the previous state H_t and the current estimate $\tilde{H}$. Finally, the maximum and the average of the output of all timesteps are concatenated and fed to a softmax classifier.

4. EXPERIMENTS

4.1. Experimental setting

Dataset In our experiments, we used the GeorgiaTech Social Interaction Dataset [2], consisting of 42 hours of video recorded by 8 different subjects wearing a head-mounted GoPro camera in a Disney World Resort. In total there are 1141 annotated video clips, with five types of interaction labels: dialogue, discussion, monologue, walk dialogue and walk discussion. These interactions take place at a dinner table with group of friends, while walking, or while standing in a line, in public transport etc. We stress that this is the only available video dataset with conversation type's annotations. Therefore, we further annotated the EgoSocialInteraction dataset [5], including 693 sequences captured by a NarrativeClip worn as a necklace, with the same labels.

Evaluation metrics Given the unbalanced distribution of labels in the dataset, in addition to the top-1 accuracy, we used also the F-score and we plot the confusion matrix.

Implementation details In the GeorgiaTech dataset we filtered out individuals with a low level of presence in video shots by computing the percentage of frames in which each person appears in a video shot weighted by the average distance from camera wearer. For the remaining people, when a person moves out of the scene for a few frames, we assume that his/her features could be inferred by interpolation from their corresponding features in temporal adjacent frames. As in [2], we trained our model on five users and tested on three. In EgoSocialRelation dataset we found only up to three instance of the *walk* classes and *monologue*. In fact, being the camera worn on the chest instead of the head, people walking side to side are hardly visible in the pictures. Therefore we did not consider these classes in our experiments.

For both datasets, we performed data augmentation directly on the feature vectors by adding random noise in the direction of the eigenvectors and proportional to the eigenvalues of the feature matrix, multiplied by a Gaussian random variable with zero mean and standard deviation 10^{-4} [6]. We employed a R-GCN with a single layer. Adding layers did not result in further improvements. We performed a grid search over the network's hyperparameters by training multiple models and choosing the one with best performance on the validation set. The best results were achieved after 83 and 32 epochs for the GeorgiaTech and the EgoSocialRelation datasets respectively, by using a learning rate of 10^{-6} , a weight decay of 0.005, and employing the Adam optimizer with both L_1 and L_2 regularization.

Comparative results In Tab. 1, we compare the proposed approach to several baseline methods. The first baseline is the method [2], that takes as input the same features employed as input to our InteractionGCN model, except the histogram of roles that was computed as in [2], but relying on the pattern of attention computed as in our approach. We

Model	GeorgiaTech		EgoSocialRelation	
	Acc	F-score	Acc	F-score
HCRF [2]	38.99%	38.99%	55.55%	63.64%
MLP	44.03%	27.27%	64.72%	63.41%
T-GCN [28]	45.67%	44.23%	81.81%	81.81%
Ours	**61.88%**	**62.36%**	**86.36%**	**84.21%**

Table 1. Performance comparisons.

Classes	Discus.	Dial.	W. Disc.	W. Dial.	Monol.	All	
#videos	527	421	154	255	74	1431	
Removed cue	**Per class accuracy**					**Accuracy**	**F-score**
H. localization	62.5%	68.6%	58.6%	63.1%	17.6%	62.47%	62.45%
H. orientation	62.5%	68.6%	55.2%	**66.2%**	23.5%	**62.47%**	**62.47%**
M. attention	61.3%	60.8%	48.3%	**66.2%**	35.3%	59.84%	60.34%
P. distance	**63.1%**	**70.6%**	**65.5%**	55.4%	23.5%	62.20%	62.53%
F. p. motion	57.7%	52.0%	10.3%	40.0%	**41.2%**	49.71%	48.89%
All cues	57.7%	68.6%	62.1%	**66.2%**	**41.2%**	61.88%	62.36%

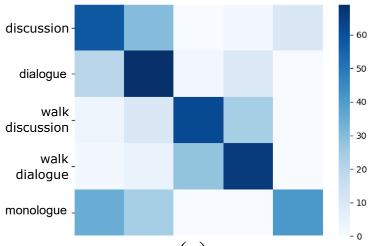
Table 2. Performance by removing an interaction cue at time.

used their HCRF model for classification[3] and report the best results obtained with 10 hidden states. In addition, we trained a Multi Layer Perception (MLP) Network, since it has proved to be significantly better than linear models for video classification. Furthermore, we implemented the Temporal Graph Convolutional Network (TGCN) model presented in [28], that connects temporally corresponding nodes but using a RGCN instead of simpler GCN. These experiments show that our model achieves a significant improvement over all baseline on both datasets.In the EgoSocialRelationDataset we can observe the same performance ranking but with overall better numerical results due the lower number of classes founded in this dataset.

4.2. Results

Ablation study on the architecture To validate the different components of InteractionGCN architecture, we report on Tab.1 (b) results obtained 1) by using GRU architecture taking as input the extracted features, and 2) by using R-GCN model to get the conversational context at the frame level, that we stacked with first-person motion features to get the class predictions, without GRU. These results, obtained on the GeorgiaTech dataset, validate all components of our framework.

Ablation study on the features Tab. 2 shows the effect of each feature employed in our model. A large performance drop is experienced when first-person motion is removed. Indeed, first-person motion is specially important to characterize walk classes. However, even without using this feature, we outperform all baselines. Instead, neglecting mutual attention as cue, specially degrades the classes *monologue* and *walk discussion*. When removing head localization or head orientation or pairwise distance, the accuracy of the class *monologue* becomes lower than or very close to the chance level, which is totally unsatisfactory for a classification problem with five classes. Instead, by keeping all the features, we overall lose less than 1% globally and up to 5% for the class *discussion*, hence obtaining a more balanced per class accuracy. Indeed, during a *monologue* most of people are looking at the person who is talking and at a distance typically larger that the one that characterizes other interactions.



(a)

Model	Acc.	F-score
GRU	53.02%	53.63%
R-GCN	48.56%	47.56%
Ours	**61.88%**	**62.36%**

(b)

Fig. 3. (a) Confusion matrix and (b) ablation study.

Discussion Fig. 3 shows the confusion matrix obtained with our InteractionGCN framework by using all features. Similarly to what was observed in [2], and to a major extent in this work, walk classes are well discriminated from non-walk classes. Contrary to what reported in [2], where (walk) dialogue and (walk) discussion are more significantly confused, in our model the class monologue is the one that is more often misclassified. This is likely due to the fact that our model of attention relies on detected faces in a video clip, whereas the one used in [2] does not. Due to the unconstrained nature of the videos, face detection may fail especially when people are wearing accessories such as hats or glasses, hence making this model more prone to erros. Additionally, *monologue* is a class with a considerable smaller number of examples with respect to the others in the dataset.

5. CONCLUSIONS

We presented InteractionGCN, a GCN based framework for categorising social interactions based on the interactivity level in sequences captured by a wearable camera. InteractionGCN extracts and models relational and non-relational interaction cues at the frame level through a graph, where people are represented by nodes and pairwise relations between people are expressed through edges. 3D head orientation and 2D localization are employed as node features, while pairwise distance and mutual attention are modeled as edge relations. It first captures the interaction context by a R-GCN on the graph at the frame level, and then feeds this information to a GRU-based architecture, together with first-person motion information. Experimental results on two public datasets demonstrated the benefits of InteractionGCN over several baselines.

[3] https://github.com/yalesong/hCRF-light

6. REFERENCES

[1] M. Dimiccoli, "Computer vision for egocentric (first-person) vision," in *Computer Vision for Assistive Healthcare*, pp. 183–210. Elsevier, 2018.

[2] A. Fathi, J. K. Hodgins, and J. M. Rehg, "Social interactions: A first-person perspective," in *CVPR*. IEEE, 2012, pp. 1226–1233.

[3] S. Bano, T. Suveges, J. Zhang, and S. J. Mckenna, "Multimodal egocentric analysis of focused interactions," *IEEE Access*, vol. 6, pp. 37493–37505, 2018.

[4] M. Aghaei, M. Dimiccoli, and P. Radeva, "With whom do I interact? detecting social interactions in egocentric photo-streams," in *ICPR*. IEEE, 2016, pp. 2959–2964.

[5] E. Sánchez-Aimar, P. Radeva, and M. Dimiccoli, "Social relation recognition in egocentric photostreams," in *ICIP*. IEEE, 2019, pp. 3227–3231.

[6] M. Aghaei, M. Dimiccoli, C. Canto-Ferrer, and P. Radeva, "Towards social pattern characterization in egocentric photo-streams," *CVIU*, vol. 171, pp. 104–117, 2018.

[7] K. Cho and et al., "Learning phrase representations using rnn encoder–decoder for statistical machine translation," 2014.

[8] J. Li, Y. Wong, Q. Zhao, and M. S. Kankanhalli, "Dual-glance model for deciphering social relationships," in *ICCV*, 2017, pp. 2650–2659.

[9] G. Wang, A. Gallagher, J. Luo, and D. Forsyth, "Seeing people in social context: Recognizing people and social relationships," in *ECCV*. Springer, 2010, pp. 169–182.

[10] Q. Dai, P. Carr, L. Sigal, and D. Hoiem, "Family member identification from photo collections," in *WACV*. IEEE, 2015, pp. 982–989.

[11] S. Xia, M. Shao, J. Luo, and Y. Fu, "Understanding kin relationships in a photo," *TMM*, vol. 14, no. 4, pp. 1046–1056, 2012.

[12] Y. Guo, H. Dibeklioglu, and L. van der Maaten, "Graph-based kinship recognition," in *ICPR*. IEEE, 2014, pp. 4287–4292.

[13] Y.-Y. Chen, W. H. Hsu, and H.-Y. M. Liao, "Discovering informative social subgraphs and predicting pairwise relationships from group photos," in *ACM MM*, 2012, pp. 669–678.

[14] Q. Sun, B. Schiele, and M. Fritz, "A domain based approach to social relation recognition," in *CVPR*, 2017, pp. 3481–3490.

[15] V. Ramanathan, B. Yao, and L. Fei-Fei, "Social role discovery in human events," in *CVPR*, 2013, pp. 2475–2482.

[16] T. Lan, L. Sigal, and G. Mori, "Social roles in hierarchical models for human activity recognition," in *CVPR*. IEEE, 2012, pp. 1354–1361.

[17] T. Shu, D. Xie, B. Rothrock, S. Todorovic, and S. Chun Zhu, "Joint inference of groups, events and human roles in aerial videos," in *CVPR*, 2015, pp. 4576–4584.

[18] X. Liu, W. Liu, M. Zhang, J. Chen, L. Gao, C. Yan, and T. Mei, "Social relation recognition from videos via multi-scale spatial-temporal reasoning," in *CVPR*, 2019, pp. 3566–3574.

[19] S. Alletto, G. Serra, S. Calderara, and R. Cucchiara, "Understanding social relationships in egocentric vision," *PR*, vol. 48, no. 12, pp. 4082–4096, 2015.

[20] H. S. Park, E. Jain, and Y. Sheikh, "3d social saliency from head-mounted cameras," in *NIPS*, 2012, pp. 431–439.

[21] H. Park and J. Shi, "Social saliency prediction," in *CVPR*. IEEE, 2015, pp. 4777–4785.

[22] M. Aghaei, M. Dimiccoli, and P. Radeva, "Towards social interaction detection in egocentric photo-streams," in *ICMV*, 2015, pp. 987514–987519.

[23] M. Aghaei, M. Dimiccoli, and P. Radeva, "All the people around me: face discovery in egocentric photo-streams," in *ICIP*. IEEE, 2017, pp. 1342–1346.

[24] X. Qi, R. Liao, J. Jia, S. Fidler, and R. Urtasun, "3d graph neural networks for rgbd semantic segmentation," in *ICCV*, 2017, pp. 5199–5208.

[25] X. Wang and A. Gupta, "Videos as space-time region graphs," in *ECCV*, 2018, pp. 399–417.

[26] H. Jiang and K. Grauman, "Seeing invisible poses: Estimating 3d body pose from egocentric video," in *CVPR*. IEEE, 2017, pp. 3501–3509.

[27] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *ESWC*, 2018, pp. 593–607.

[28] S. Yan, Y. Xiong, and D. Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in *AAAI*, 2018.