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 Interac-
tionGCN. Our method extracts patterns of relational and non-
relational 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, to-
gether with first-person motion information, through a Gated
Recurrent Unit architecture. Ablation studies and experimen-
tal evaluation on two publicly available datasets validate the
proposed approach and establish state of the art results.

Index Terms— social interaction recognition, graph con-
volutional networks, egocentric vision

1. INTRODUCTION
Recently, wearable cameras have enabled the automatic cap-
ture 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 detec-
tion [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 (dis-
cussion 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

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

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

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 [2], captured by
a head-mounted wearable camera in an amusement park: five
different conversation types may occur at different time in-
tervals during a social interaction, even in absence of drastic
visual changes, e.g. discussion vs dialogue. In [2], this prob-
lem 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.

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 associ-
ated individual features, and edges correspond to interaction
cues between pairs of persons. We learn the interactivity con-
text 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) together
with first-person motion. A visual overview of our Interac-
tionGCN framework is given in Fig. 2. Experimental results
on two publicly available datasets validate the proposed
approach.

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

1http://cbs.ic.gatech.edu/egocentric/datasets.htm
in personal photo albums \cite{10,13}. More recently, a formalization of people social life based on the Bugental theory has been proposed \cite{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 \cite{15,17} by relying on relative age, gender, clothing and again on people relative location. Recently, \cite{18} proposed a Multi-scale Spatial-Temporal Reasoning framework to classify social interactions from videos into the eight more common sub-categories among the ones proposed in \cite{14}.

**First-person social interaction recognition** So far most efforts have focused on detecting groups of interacting people \cite{19}, detecting social saliency \cite{20,21}, or detecting with whom the camera wearer is interacting \cite{4,22}. Only a few works have gone beyond the detection task in the egocentric domain \cite{5,6,23}. Aghaei et al. \cite{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 \cite{5}. Similarly to us, \cite{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 \cite{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 \cite{2,9,10,12,15,16} have been little explored in this context so far \cite{18}. However, this work focuses 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 \cite{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 \cite{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.

**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 \footnote{\url{https://azure.microsoft.com/en-us/services/cognitive-services/}}. Finally, first-person motion features were estimated by using the method proposed in \cite{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 \cite{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 \( \mathcal{G} = (V, E, R, W) \) 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), \( W \) 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 \( (e_{ij} \in E, r \in R = \{d, a\}) \), corresponding to two different pairwise relations: relative distance between pairs of individuals.
in the scene \( (d) \), including the camera wearer, and directional attention pattern \( (a) \). Distance edges \( (e_{ij}^d) \) are bi-directional and represent the distance between \( P_i \) and \( P_j \), while attention edges \( (e_{ij}^a) \) are directed edges pointing from \( P_i \) to \( P_j \), meaning that \( P_i \) is looking at \( P_j \). \( \alpha_{ij}^a \in \mathcal{W} \) is the weight of the labeled edge \( e_{ij}^a \). 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:

\[
h_{i}^{(k+1)} = \sigma \left( \sum_{r \in \mathcal{R}} \sum_{j \in N_i^r} \alpha_{ij}^r W_r h_{j}^{(k)} + \alpha_{0i}^r W_0 h_{i}^{(k)} \right)
\]

where \( h^{(k+1)} \) is the hidden state of node \( v_i \) at the \( k + 1 \) layer in a R-GCN, \( \sigma \) is the ReLU activation function, \( W_r \) and \( W_0 \) are learnable parameters of the transformation, \( \alpha_{ij}^r \) and \( \alpha_{0i}^r \) are the edge weights, \( N_i^r \) 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 \( c_{i,r} = |N_i^r| \). 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:

\[
z_t = \sigma(W_{xz} \cdot x_t + W_{Hx} \cdot H_t + b_z)
\]

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

\[
\tilde{H} = \tanh(W_{xH} \cdot x_t + W_{HH} \cdot (r_t \ast H_t) + b_H)
\]

\[
H_{t+1} = z_t \ast H_t + (1 - z_t) \ast \tilde{H},
\]

where \( x_t \) and \( H_{t+1} \) are the input and output vectors, \( z_t \) and \( r_t \) are the update and reset vectors, \( W_z \) and \( b_z \) 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 Narrative-Clip 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
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.

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 errors. 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.
6. REFERENCES


