SOCIAL RELATION RECOGNITION IN EGOCENTRIC PHOTOSTREAMS

Emanuel Sánchez Aimar(1), Petia Radeva(1), Mariella Dimiccoli(2)

(1) University of Barcelona, Computer Vision Center, Barcelona, Spain.
(2) Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain

ABSTRACT

This paper proposes an approach to automatically categorize the social interactions of a user wearing a photo-camera (2fm), by relying solely on what the camera is seeing. The problem is challenging due to the overwhelming complexity of social life and the extreme intra-class variability of social interactions captured under unconstrained conditions. We adopt the formalization proposed in Bugental’s social theory, that groups human relations into five social domains with related categories. Our method is a new deep learning architecture that exploits the hierarchical structure of the label space and relies on a set of social attributes estimated at frame level to provide a semantic representation of social interactions. Experimental results on the new EgoSocialRelation dataset demonstrate the effectiveness of our proposal.

Index Terms— social relation recognition, egocentric vision, multi-task learning, LSTM

1. INTRODUCTION

As our social life keeps moving towards the digital world and its social networks, new collective moments are continuously being captured in the form of pictures, audio, videos, and text. Meanwhile, several studies have shown that human relationships have an important effect in human health, involving physical and mental health, behaviour, and mortality risk [1]. The need of a broader understanding of our social relations and their influence on human health have motivated an increasing interest in the computer vision community for automatic discovery, quantification and categorization of social interactions from the vast amount of public images and videos [2, 3, 4, 5, 6, 7]. Recently, Aghaei et al. [7] have shown the usefulness of egocentric photostreams, captured by a wearable photo-camera [8] to automatically analyze the daily social interactions of a person, in a natural setting where people appear in an intimate perspective. Despite the challenging characteristics of the egocentric domain, such as the fact that the user is not visible in the field of view, background clutter, and abrupt appearance changes [7], the authors showed that it is possible not only to understand when the camera wearer is interacting with somebody, but also to determine with how many people the user has interacted with during a given period of time, the duration and frequency of the interactions. However, in [7] the classification of interactions was limited to formal and informal meetings. Recent work [6] has proposed the Bugental’s domain-based social theory [9] as a conceptualization of human social life to categorize social interactions in images. The theory includes five domains with examples of common relations, characterized by specifics attributes and behaviours. Nevertheless, in [6], this approach has been applied on a dataset of third-person images collected from photo albums.

To evaluate the formalization proposed in [6] in a naturalistic setting and at the same time going deeper into the understanding of social interactions from egocentric photostreams, in this work we propose a new egocentric dataset, hereafter referred to as EgoSocialRelation dataset [1], where social interactions, formed of short image sequences, are annotated in a hierarchy of domain and relation labels, derived from Bugental’s theory. Furthermore, we propose and validate, for the first time on egocentric data, several models for social relation categorization, hence providing a solid benchmark for further studies. The proposed models rely on the composition of visual semantic attributes, and exploit the sequential nature of photostreams. Code for experiments is publicly released.

The rest of the paper is as follows: section 2 reviews related work; section 3 details our approach; section 4 introduces a new dataset and discusses experimental results. Section 5 concludes this work summarizing its contributions.

2. RELATED WORK

Third-view domain A large body of work has focused on family member recognition [3], role recognition in so-

Fig. 1: Examples of images from EgoSocialRelation dataset.
and categorized interactions into two broad categories, namely w.r.t. the camera-user. Extending the previous work, [7] while exploiting distance and orientation of observable people, they proposed to detect social interactions in egocentric photostreams, that is currently limited to formal/informal meeting interactions from egocentric photostreams, which go beyond the state of the art on the classification of social interactions.

With this work we mark methods to classify social interactions from egocentric videos, while taking into account for contextual information. They argue that for effective detection and categorization of social interactions from an egocentric perspective, a combination of social signals and environmental features is needed, as well as their evolution over time.

In this paper, we propose a new dataset and several benchmark methods to classify social interactions from egocentric photostreams into five domains and nine relations, following the conceptualization of Bugental’s theory. With this work we go beyond the state of the art on the classification of social interactions from egocentric photostreams, that is currently limited to formal/informal meeting classification.

3. METHODOLOGY

We propose several deep learning architecture that leverage multiple semantic attributes and their temporal evolution over time. In particular, we aim at investigating the importance of semantic attributes for the classification performance in egocentric photostreams as well as how to take advantage of the hierarchical label space.

3.1. Preprocessing

Each photostream is partitioned into semantically meaningful segments by applying SRelustering [15], where segments with a high ratio of visible people relative to the number of frames are considered as social segments. Given a frame in a social segment, we extract three different regions, illustrated in fig. 2. First, we apply a face detector [16] to extract visible faces, discarding candidates with confidence score (IoU) below 0.99. Then, we create an initial estimate of face clusters based on visual similarity, using Microsoft Cognitive Service API. It follows a manual procedure based on visual inspection, to improve the quality of the clustering (recategorizing misclassified samples, adding uncategorized ones, discarding spurious samples and creating new clusters if needed). For each observable person, we reorganize sub-segments with valid faces. Given a valid frame, we extract Face and Body regions, the latter delimited by 3 x face width and 6 x face height, inspired in [6, 17]. Finally, we denote the full (original) image as Contextual region. The procedure results in a new dataset of sequences with trackable people hereafter referred to as user-specific segments.

3.2. Feature extraction

We leverage CNN models pretrained on specialized datasets to predict human-related attributes. Given a user-specific segment, for each frame and social cue, we extract high-dimensional intermediate CNN features, aka visual embeddings. We remove the task-specific classification layer, ultimately using the penultimate fully connected (FC) layer.

Table 1 lists the semantic attributes used in this work. In addition, because it is not possible to observe the person holding the camera, we include the camera-wearer’s ground truth age and gender information, following the categorization in [6]. If we were to concatenate all extracted features, the global representation would add up 33801 variables. To mitigate the curse of dimensionality phenomenon, we apply dimensionality reduction to CNN features for each attribute independently, based on the approach proposed by [7]. In this work, we define the quantification factor \( Q = 32 \), while keeping the 50 most relevant principal components (ensuring enough level of detail, with explained variance around 90%). After merging all the semantic attributes, including compressed CNN features along with no-CNN features, we obtain a final representation with 459 variables.

![Fig. 2: Illustration of face, body and contextual regions.](image-url)
4. EXPERIMENTAL RESULTS

4.1. Experimental setting

Dataset We started from the EgoSocialStyle dataset, collected by 9 users wearing a Narrative Clip camera, recording at two fps in a daily life scenario. Following the protocol in [7], we extended it with 119 new sequences, collected by 9 users wearing a Narrative Clip camera, recording at two fps in a daily life scenario. We started from the Dataset (grouped in five domains and nine relations, namely Section 3.1) and annotated them with social labels. Later on, we extracted user-specific segments inspired by the hierarchical approach proposed by Cerri et al. [24]. We also evaluate the model in fig. 3 without the extra constrain (denoted as the Multi-task Independent strategy, (MT-IND)). As the model’s description suggests, multiple objectives are trained with multi-task learning [25], jointly optimizing the loss functions (with equal importance). In all cases, the first and last FC layers are followed by ReLU and Softmax activation functions respectively, while using Cross-Entropy as loss function.

Validation methodology To validate our approach, we used a form of repeated random sub-sampling cross-validation [26]. First, we arrange our dataset in groups of whole days captured by a given user. Then, we sampled randomly \( N = 1000 \) examples, ensuring the day’s separation criteria and approximately 80\%/20\% size ratios for training and validation, respectively. For each combination, we considered the best candidates with minimal Kullback-Leibler divergence between the normalized distributions of each split. We pick the top candidate, leaving the validation split for testing purposes, and repeat the adhoc procedure using the training split, to obtain the top \( K=3 \) splits for model cross-validation. This strategy allows us to define data splits in a way that overlapping or consecutive user-specific sequences, that may capture the same social interaction, are put together, while maintaining the statistical distribution of the data. Given the relatively small size of our dataset, we applied the data augmentation strategy proposed by [7] to mitigate the over-fitting problem. In a glance, we compute PCA and add random noise in the direction of the eigenvectors, and proportional to the eigenvalues times a Gaussian random variable \( X \sim N(\mu = 0, \sigma = 0.01) \). This way we ensure that the original labels are preserved in new augmented samples. Since the dataset is highly imbalanced, we assess the competing models with two metrics, overall accuracy (abbreviated acc) and macro f1-score. We maximize f1-score for model selection, giving the same importance to all classes, instead of performing well just on over-represented classes. We address class imbalance further by using a class weighting scheme embedded in the global loss function [27]. It follows our final model configuration, obtained with grid search over the next parameters: number of neurons \( = 128 \), learning rate \( \alpha = 2e^{−3} \), dropout rate \( = 0.3 \), L2 regularization \( \lambda = 1e^{−3} \) and number of training iterations \( = 150 \). We used Adam [28] to optimize the model, with a step decay schedule halving the initial \( \alpha \) every 50 iterations.

4.2. Discussion

Social relations Although f1-score and acc validate the results in distinct ways, both follow the same trend for models categorizing relations in table 2 (prefix REL and DOM for...
relation and domain recognition, respectively). By leveraging the hierarchy of labels and injecting knowledge of domains, model REL-MT-TD achieves the highest performance in relation recognition, this way proving beneficial to have both coarse and fine-grained social categorizations. This extra hint is key to our approach, as training with independent objectives (REL-MT-IND) seems counterproductive, even compared to not exploiting domain labels at all (REL-ST). For comparative purposes, we observe that Sun et al. [6], reported an acc equivalent to REL-ST for relation recognition on PIPA (REL-SVM-PIPA, f1-score not available).

Social domains Social characterization at domain level provides useful information, despite the coarser representation. Table 2 presents single-task model DOM-ST as the top performer for this task, surpassing alternatives that rely on relation labels, what indicates that a finer class granularity does not necessarily improve predictions at the top level. Still, model DOM-MT-IND has a f1-score 10% lower that DOM-MT-TD, further supporting the top-down approach. Given the hierarchical label space, we can either predict the most likely domain as presented before, or indirectly predicting the most likely relation, and then inferring the associated domain. Notoriously, the performance of multi-task strategies increases with the second approach, model DOM-MT-TD a boost in f1-score and acc up to 42.69% and 59.40% respectively, matching the best model’s accuracy. In comparison, Sun et al. [6] reports a slightly higher acc of 67.80% in PIPA, possibly due to a difference in criteria for model selection (we used f1-score instead of acc).

Analysis of semantic attributes In this section, we study the contribution of different subsets of semantic attributes (see table 3), with focus on the MT-TD strategy to simplify the analysis. Models FACE and BODY include facial and body attributes (respectively), and extra camera-user’s info.

![Fig. 4: F1-score per domain class by group of attributes.](image)

Model CTX exploits the context by considering activity and proximity, while model ALL denotes the fusion of all attributes. We observe that the contribution of partial subsets of attributes is stronger for domain recognition, or equivalently, more attributes are needed to recognize relations. This support the hypothesis considering the relation recognizing as a more complex, most likely due to finer class granularity. Nevertheless, both tasks maximize f1-score by leveraging all attributes. It can be seen that BODY models present considerable higher acc than their FACE counterparts, most likely due to body features being more robust to partial occlusion and different perspectives. However, this fact contrasts with f1-score performance. To shed light on this issue, fig. 4 presents f1-score computed for each domain class.

Facial attributes are specially relevant for Attachment and Mating (emotions, head orientation, gender cue). Furthermore, without facial information, f1-score drops heavily for Mating, acknowledging that lovers may not be distinguished from friends or co-workers in this scenario. In line with previous studies [7], faces are key to categorize colleagues (Coalitional group) and friends (Reciprocity). Body attributes provide a different perspective, still, they are very relevant for Coalitional group (e.g. uniform clothing) and Reciprocity, and also for Attachment, characterized by large age difference. Finally, daily activities (main contextual signal) are key to classify Coalitional group (working) and Reciprocity (gathering and sharing). Summarizing, in our experiments we observe that social domains respond to specific social cues, in correspondence with the principles proposed by Bugental.

### Table 2: Social relation and domain recognition results.

| Relation | F1-score [%] | Acc [%] |
|----------|--------------|---------|
| REL-FACE | 23.39        | 31.60   |
| REL-BODY | 25.30        | 49.60   |
| REL-CTX  | 25.18        | 46.60   |
| REL-ALL  | 33.26        | 58.60   |
| DOM-FACE | 34.42        | 38.30   |
| DOM-BODY | 31.69        | 50.40   |
| DOM-CTX  | 33.61        | 45.90   |
| DOM-ALL  | 42.49        | 56.40   |

Table 3: Recognition results by attribute groups with MT-TD.

### Table 3: Recognition results by attribute groups with MT-TD.

| Attribute | F1-score [%] | Acc [%] |
|-----------|--------------|---------|
| FACE      | 23.39        | 31.60   |
| BODY      | 25.30        | 49.60   |
| CTX       | 25.18        | 46.60   |
| ALL       | 33.26        | 58.60   |
| FACE      | 34.42        | 38.30   |
| BODY      | 31.69        | 50.40   |
| CTX       | 33.61        | 45.90   |
| ALL       | 42.49        | 56.40   |

5. CONCLUSIONS

This paper addressed for the first time the categorization of social relations following Bugental’s conceptualization in the domain of egocentric photostreams. A new egocentric dataset of social events acquired under unconstrained conditions, has been released and a family of models employing CNN models for feature extraction and a LSTM-based classifier have been tested providing a benchmark. Moreover, by applying multi-task learning with a hierarchical label space in a top-down approach, our model provides a solid baseline for the task of relation recognition, while outperforming straightforward alternatives.
