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.mdimiccoli@iri.upc.edu

ABSTRACT

This paper proposes an approach to automatically categorize the social interactions of a user wearing a photo-camera (2fpm), 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 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.

Fig. 1: Examples of images from *EgoSocialRelation* dataset.

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[2].

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-

[1] https://chest.iri.upc.edu/files/users/mdimiccoli/public_html/DATASETS/EgoSocialRelation.zip
[2] https://github.com/emasa/social-relations-recognition-egocentric-photostreams

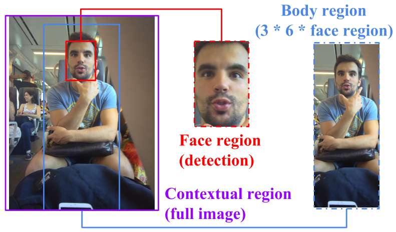

Fig. 2: Illustration of face, body and contextual regions.

cial events [2], categorization of social groups based on appearance (e.g. urban tribes [4]) and occupation recognition based on contextual information [5]. Inspired on Bugental's social domain-based theory [9], the authors of [6] proposed an *holistic* approach to social relations categorization from image data. The published dataset contains *still images* with labels in a hierarchical structure of 5 social domains and 16 social relations, derived from Bugental's theory. Furthermore, they derived a family of models, combining *human attributes* and *behaviours*, extracted with pretrained Convolutional Neural Networks (CNN), to ultimately classify the CNN feature representation with a Support Vector Machine (SVM).

Egocentric domain The seminal work of Fathi *et al.* [10] proposed to detect social interactions from egocentric videos, by inferring the 3D location to which a person is looking at, through a Markov Random Field model. They further categorized these interactions into three classes, namely *discussion*, *dialogue* and *monologue*, depending on the role of the participants in the interaction. Later on, [11] proposed a procedure to analyze social interaction sequences from egocentric videos and detect them applying a model based on Hidden Markov Model and SVM, focusing on head and body poses. With a different approach, Aghaei *et al.* [12, 13] proposed to detect social interactions in egocentric photostreams, while exploiting distance and orientation of observable people w.r.t. the camera-user. Extending the previous work, [7] categorized interactions into two broad categories, namely *formal* and *informal meetings*, based on the classification of sequences with a Long Short Term Memory (LSTM) model [14]. The authors extended the original attributes with facial expressions and global CNN features extracted at frame level, the former related to affective internal states and the latter to 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 *SRclustering* [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 subsegments 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.

Semantic attribute	CNN architecture	Output layer	Image region	Dataset	Source
(Daily) Activities	ResNet50 [18]	*relu5c*	Full	[19]	[19]
Age (x2)	CaffeNet [20]	*relu7*	Face Body	PIPA [17]	[6]
Clothing	CaffeNet [20]	*relu7*	Body	Berkeley People Attributes [21]	[6]
Facial Expression	CaffeNet [20]	*relu7*	Face	IMFDB [22]	[6]
Gender (x2)	CaffeNet [20]	*relu7*	Face Body	PIPA [17]	[6]
Head Appearance	CaffeNet [20]	*relu7*	Face	CelebA [23]	[6]
Head Orientation	CaffeNet [20]	*relu7*	Face	IMFDB [22]	[6]
Proximity	N/A	N/A	Full	[7]	[7]

Table 1: List of semantic attributes.

3.3. Time-series classification

Given the compact representation computed for each frame in a *user-specific segment*, we pose the problem of social relation recognition as multi-class time-series classification, where each component along the time axis represents the time-evolution of an individual feature. We employ a LSTM [14], specially designed to learn long-term dependencies in the sequence. Our simplest model architecture is presented in fig. 3a (denoted as *Single-task* strategy (**ST**)). Two different models are trained using either relation or domain labels. To take advantage of the hierarchical label space, we propose the model in fig. 3b. The class output of the first level, i.e. domains, is provided as input for predicting the second level, i.e. relations (denoted as *Multi-task Top-down strategy* (**MT-TD**)), inspired on the hierarchical approach proposed by Cerri *et al.* [24]. We also evaluate the model in fig. 3c 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.

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 fpm in a daily life scenario. Following the protocol in [7], we extended it with 119 new sequences, collected by the same users. Later on, we extracted *user-specific segments* (section 3.1) and annotated them with social labels. Similar to [6], valid sequences must have **one valid relation label**. We discarded examples with zero or multiple labels, also with insufficient samples (less than 20). Our final *EgoSocialRelation* dataset includes 693 sequences, grouped in five domains and nine relations, namely *Attachment* (*father-child*, *mother-child*), *Reciprocity* (*friends*, *classmates*), *Mating* (*lovers*), *Coalitional group* (*colleagues*), and *Hierarchical group* (*presenter-audience*, *leader-subordinate* and *customer-staff*).

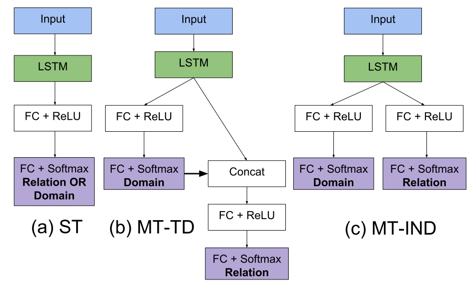

Fig. 3: Evaluated architectures for social recognition.

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 overfitting 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 $X \sim \mathcal{N}(\mu = 0, \sigma = 0.01)$. 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 $\alpha = 2\text{e}{-3}$, dropout rate $= 0.3$, L2 regularization $\lambda = 1\text{e}{-3}$ and number of training iterations $= 150$. We used Adam [28] to optimize the model, with a *step decay schedule* halving the initial α 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

	F1-score [%]	Acc [%]
REL-ONLY	32.19	57.10
REL-MT-I	31.06	54.90
REL-MT-TD	**33.26**	**58.60**
DOM-ONLY	**44.52**	**59.40**
DOM-MT-I	38.38	54.90
DOM-MT-TD	42.49	56.40
REL-SVM-PIPA	-	57.20
DOM-SVM-PIPA	-	67.80

Table 2: Social relation and domain recognition results.

	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.

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, giving 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.

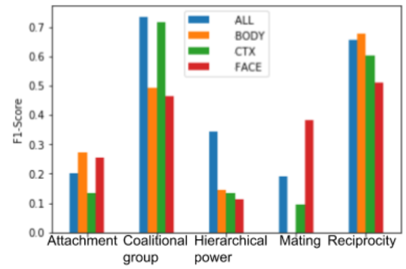

Fig. 4: F1-score per domain class by group of attributes.

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.

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.

6. REFERENCES

[1] Debra Umberson and Jennifer Karas Montez, "Social relationships and health: A flashpoint for health policy," *Journal of health and social behavior*, vol. 51, no. 1_suppl, pp. S54–S66, 2010.

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

[3] Afshin Dehghan, Enrique G Ortiz, Ruben Villegas, and Mubarak Shah, "Who do i look like? determining parent-offspring resemblance via gated autoencoders," in *IEEE CVPR*, 2014, pp. 1757–1764.

[4] Ana C Murillo, Iljung S Kwak, Lubomir Bourdev, David Kriegman, and Serge Belongie, "Urban tribes: Analyzing group photos from a social perspective," in *CVPRW*. IEEE, 2012, pp. 28–35.

[5] Ming Shao, Liangyue Li, and Yun Fu, "What do you do? occupation recognition in a photo via social context," in *ICCV IEEE*, 2013, pp. 3631–3638.

[6] Qianru Sun, Bernt Schiele, and Mario Fritz, "A domain based approach to social relation recognition," in *IEEE CVPR*, 2017, pp. 21–26.

[7] Maedeh Aghaei, Mariella Dimiccoli, Cristian Canton Ferrer, and Petia Radeva, "Towards social pattern characterization in egocentric photo-streams," *Computer Vision and Image Understanding*, vol. 171, pp. 104–117, 2018.

[8] Marc Bolanos, Mariella Dimiccoli, and Petia Radeva, "Toward storytelling from visual lifelogging: An overview," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 1, pp. 77–90, 2017.

[9] Daphne Blunt Bugental, "Acquisition of the algorithms of social life: a domain-based approach.," *Psychological bulletin*, vol. 126 2, pp. 187–219, 2000.

[10] Alireza Fathi, Jessica Hodgins, and James Rehg, "Social interactions: A first-person perspective," pp. 1226–1233, 06 2012.

[11] Jen-An Yang, Chia-Han Lee, Shao-Wen Yang, V Srinivasa Somayazulu, Yen-Kuang Chen, and Shao-Yi Chien, "Wearable social camera: Egocentric video summarization for social interaction," in *ICMEW IEEE*, 2016, pp. 1–6.

[12] Maedeh Aghaei, Mariella Dimiccoli, and Petia Radeva, "Towards social interaction detection in egocentric photo-streams," in *ICMV*, 2015, vol. 9875.

[13] Maedeh Aghaei, Mariella Dimiccoli, and Petia Radeva, "With whom do i interact? detecting social interactions in egocentric photo-streams," in *ICPR 2016*. IEEE, 2016, pp. 2959–2964.

[14] Sepp Hochreiter and Jürgen Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] Mariella Dimiccoli, Marc Bolaños, Estefania Talavera, Maedeh Aghaei, Stavri G Nikolov, and Petia Radeva, "Sr-clustering: Semantic regularized clustering for egocentric photo streams segmentation," *Computer Vision and Image Understanding*, vol. 155, pp. 55–69, 2017.

[16] N. Ruiz and J. M. Rehg, "Dockerface: an easy to install and use Faster R-CNN face detector in a Docker container," *ArXiv e-prints*, Aug. 2017.

[17] Seong Joon Oh, Rodrigo Benenson, Mario Fritz, and Bernt Schiele, "Person recognition in personal photo collections," in *ICCV IEEE*, 2015, pp. 3862–3870.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE CVPR*, 2016, pp. 770–778.

[19] Alejandro Cartas, Juan Marín, Petia Radeva, and Mariella Dimiccoli, "Batch-based activity recognition from egocentric photo-streams revisited," *Pattern Analysis and Applications*, vol. 21, no. 4, pp. 953–965, 2018.

[20] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *ACM Multimedia*, 2014, pp. 675–678.

[21] Lubomir Bourdev, Subhransu Maji, and Jitendra Malik, "Describing people: A poselet-based approach to attribute classification," in *IEEE ICCV*, 2012, pp. 1543–1550.

[22] Shankar Setty, Moula Husain, Parisa Beham, Jyothi Gudavalli, Menaka Kandasamy, Radhesyam Vaddi, Vidyagouri Hemadri, JC Karure, Raja Raju, B Rajan, et al., "Indian movie face database: a benchmark for face recognition under wide variations," in *CVPRIPG*. IEEE, 2013, pp. 1–5.

[23] Li-Jia Li, David A. Shamma, Xiangnan Kong, Sina Jafarpour, Roelof van Zwol, and Xuanhui Wang, "Celebritynet: A social network constructed from large-scale online celebrity images," *TOMCCAP*, vol. 12, pp. 3:1–3:22, 2015.

[24] Ricardo Cerri, Rodrigo C Barros, André CPLF de Carvalho, and Yaochu Jin, "Reduction strategies for hierarchical multi-label classification in protein function prediction," *BMC bioinformatics*, vol. 17(1), pp. 373, 2016.

[25] Sebastian Ruder, "An overview of multi-task learning in deep neural networks.," *CoRR*, vol. abs/1706.05098, 2017.

[26] Ron Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *IJCAI - Volume 2*, 1995, pp. 1137–1143.

[27] Gary King and Langche Zeng, "Logistic regression in rare events data," *Political analysis*, vol. 9, no. 2, pp. 137–163, 2001.

[28] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.