Towards Social Interaction Detection in Egocentric Photo-streams

Maedeh Aghaei, Mariella Dimiccoli, Petia Radeva
University of Barcelona and Computer Vision Centre, Barcelona, Spain
aghaei.maya@gmail.com

Recent advances in wearable camera technology have led to novel applications in the field of Preventive Medicine. For some of them, such as cognitive training of elderly people by digital memories and detection of unhealthy social trends associated to neuropsychological disorders, social interaction are of special interest. Our purpose is to address this problem in the domain of egocentric photo-streams captured by a low temporal resolution wearable camera (2fpm). These cameras are suited for collecting visual information for long period of time, as required by the aforementioned applications. The major difficulties to be handled in this context are the sparsity of observations as well as the unpredictability of camera motion and attention orientation due to the fact that the camera is worn as part of clothing (see Fig. 1). Inspired by the theory of F-formation which is a pattern that people tend to follow when interacting [5], our proposed approach consists of three steps: multi-faces assignment, social signals extraction and interaction detection of the individuals with the camera wearer (see Fig. 2).

1. Multi-face Assignment

While person detection and tracking in classical videos have been active research areas for a long time, the problem of people assignment in low temporal resolution egocentric photo-streams is still unexplored. To address such an issue, we proposed a novel method for multi-face assignment in egocentric photo-streams, we called extended-Bag-of-Tracklets (eBoT) [2]. This approach basically consists of 4 major sequential modules: seed and tracklet generation, grouping tracklets into eBoT, prototypes extraction and occlusion treatment. Prior to any computation, first, a temporal segmentation algorithm [6] is applied to extract segments characterized by similar visual properties. Later on, a face detector is applied on all the frames of a segment to detect visible faces on them [8]. Based on the ratio between the number of frames with detected faces and the total number of frames of the segment, we extract segments containing trackable persons. The next steps are applied on these extracted segments, hereafter referred to as *sequences*.

Figure 1. Example of social interaction (first row) and non-social interaction (second row) in egocentric photo-streams.

- **Seed and tracklet generation:** The set of collected bounding boxes that surround the face of each person throughout the sequence, are called *seeds*. For each seed, a set of correspondences to it is generated along the sequence by propagating the seed forward and backward employing the deep-matching technique [7] that lead to form a *tracklet*. To propagate a seed found in a frame, in all the frames of the sequence, the region of the frames most similar to the seed is found as the one having the highest deep-matching score.

- **Grouping tracklets into Bag-of-tracklets (eBoT):** Assuming that tracklets generated by seeds belonging to the same person in a sequence, are likely to be similar to each other, we group them into a set of non-overlapping *eBoTs*. Since seeds corresponding to false positive detections generate unreliable tracklets and unreliable eBoTs, we defined a measure based on the *density* of the eBoTs to exclude unreliable eBoTs.

- **Prototypes extraction:** A *prototype* extracted from an eBoT, should best represent all tracklets in the eBoT, and therefore, it should best localize a person's face in each frame. As the prototype frame, the frame whose bounding box has the biggest intersection with the rest of the tracklets in that frame is chosen.

- **Occlusion treatment:** Estimation of occluded frames is a very helpful feature since it allows us to exclude occluded frames which do not convey many information from final prototypes. To this goal, we define a *frame confidence* measure to assign a confidence value

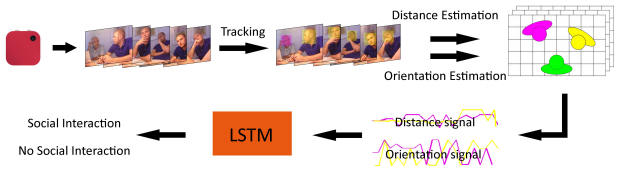

Figure 2. Outline of the proposed method for social interaction detection in egocentric photo-streams.

to each detection. When there is a severe or partial occlusion of the face, or the target is missing, the frame confidence on that frame experiences a drop, which can be robustly detected.

2. Social Signals Extraction

The F-formation model relies on a bird-view model of the scene, where each person is represented by the coordinates (x, z), so that x denotes the person position in the bird-view model and z its distance to the camera. Therefore, in our egocentric setting, we estimated the distance of people in the scene from the camera by training a polynomial regression model that, based on the camera-pinhole model, learns the relation between the distance from the camera and the vertical height of a face. The camera wearer is set on x position of the middle of the scene and on distance zero from the camera. Another important feature in the analysis of F-formations is the head orientation, which gives a rough estimation of where a person is looking at. To this goal, for visible faces, we apply the face pose estimation algorithm on the extended detected region [8]. As the camera is worn basically on the chest of the camera wearer, to predict camera wearer line of sight, we assume he can possibly look at anywhere from his left side to his right side.

3. Social Interaction Detection

To determine the potential groups of interacting people, both the pose and position vectors of each individual over all the sequence should be processed. Two different approaches for this analysis are utilized:

Hough-Voting (HVFF): The method proposed by Cristani et al. [4] for finding socially interacting groups in conventional videos is adapted to our egocentric photo-stream scenario. The method employs the Hough-voting strategy in every frame of the sequence to find the common area of interaction relying on their relative orientations and distances. By estimating pair-to-pair interaction probabilities over the sequence, the method states the presence or absence of interaction with the camera wearer and specifies which people are more involved in the interaction.

Long-Short Term Memory (LSTM): We address the problem employing particular Recurrent Neural Network (RNN) classifier, namely LSTM, to exploit its ability to use the temporal evolution of the descriptors for classification. By introducing distance and orientation feature of each person along a sequence, LSTM takes into account the temporal evolution of the features over time and is able to decide the sequence to which class (socially interacting or not) pertains. For training of LSTM, we tried two different optimization techniques: Limited memory BFGS (L-BFGS) and minibatch Stochastic Gradient Descent (SGD). We validated our method over a dataset of 20.000 images captured by Narrative camera (http://getnarrative.com/). The obtained results using both methods can be observed in Table 1 which demonstrate more accurate results using the sequence-level analysis of the social interactions using LSTM, comparing F-measures. In the future, we will employ extra features such as the facial expression evolution over time to improve the detection. Further details of our work can be found in [1, 2, 3].

	LSTM		HVFF
	LBFGS	**SGD**	
Precision	82%	73%	80%
Recall	74%	85%	72%
F-measure	77%	78%	75%

Table 1. The quantitative results of the proposed methods.

References

[1] M. Aghaei, M. Dimiccoli, and P. Radeva. Towards social interaction detection in egocentric photo-streams. In *Eighth International Conference on Machine Vision*, pages 987514–987514, 2015.

[2] M. Aghaei, M. Dimiccoli, and P. Radeva. Multi-face tracking by extended bag-of-tracklets in egocentric photo-streams. *Computer Vision and Image Understanding*, 149:146–156, 2016.

[3] M. Aghaei, M. Dimiccoli, and P. Radeva. With whom do i interact? detecting social interactions in egocentric photo-streams. *arXiv preprint arXiv:1605.04129*, 2016.

[4] M. Cristani, L. Bazzani, G. Paggetti, A. Fossati, D. Tosato, A. Del Bue, G. Menegaz, and V. Murino. Social interaction discovery by statistical analysis of f-formations. In *BMVC*, volume 2, page 4, 2011.

[5] A. Kendon. *Conducting interaction: Patterns of behavior in focused encounters*, volume 7. CUP Archive, 1990.

[6] E. Talavera, M. Dimiccoli, M. Bolaños, M. Aghaei, and P. Radeva. R-clustering for egocentric video segmentation. In *Pattern Recognition and Image Analysis*, pages 327–336. Springer, 2015.

[7] P. Weinzaepfel, J. Revaud, Z. Harchaoui, and C. Schmid. Deepflow: Large displacement optical flow with deep matching. In *ICCV*, pages 1385–1392, 2013.

[8] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, pages 2879–2886. IEEE, 2012.