Finding Time Together: Detection and Classification of Focused Interaction in Egocentric Video

Sophia Bano  Jianguo Zhang  Stephen J. McKenna
Computer Vision and Image Processing Group, School of Science and Engineering (Computing)
University of Dundee, United Kingdom
{s.bano, j.n.zhang, s.j.z.mckenna}@dundee.ac.uk

Abstract

Focused interaction occurs when co-present individuals, having mutual focus of attention, interact by establishing face-to-face engagement and direct conversation. Face-to-face engagement is often not maintained throughout the entirety of a focused interaction. In this paper, we present an online method for automatic classification of unconstrained egocentric (first-person perspective) videos into segments having no focused interaction, focused interaction when the camera wearer is stationary and focused interaction when the camera wearer is moving. We extract features from both audio and video data streams and perform temporal segmentation by using support vector machines with linear and non-linear kernels. We provide empirical evidence that fusion of visual face track scores, camera motion profile and audio voice activity scores is an effective combination for focused interaction classification.

1. Introduction

Recording of daily life experiences from a first-person perspective has become more prevalent with the increasing availability of wearable cameras used in applications such as life-logging, security, sports, ambient assisted living and driving assistance. In recent years, analysis of egocentric video has therefore gained the attention of the computer vision community. Whereas social interaction detection from a third-person perspective has been a well-researched area for some time [4, 10, 25], ego-centric vision-based methods are increasingly addressing the detection of social interaction from a first-person perspective; methods have been proposed to detect groups of individuals interacting with each other or with the camera wearer [2, 5, 13]. These methods perform off-line processing of short video clips or photo streams mostly captured from constrained perspectives and always containing interacting or non-interacting individuals. However, in reality egocentric videos are unconstrained when used for capturing daily living in long, continuous sequences.

Audio-visual feature fusion has been used for applications such as speaker localisation and event detection in social gatherings using videos captured in highly controlled indoor settings [3, 14], social interaction detection in nursing homes using surveillance-type camera videos [9], and scene change detection in life-logging videos [23]. Although audio signals provide information about social interactions, the fusion of visual and audio cues for detection of social interactions in egocentric video was rarely explored. Furthermore, the effect of integrating global camera motion analysis methods, nowadays used for human activity recognition in egocentric videos [30], with other visual and audio features for social interaction analysis still needs to be researched.

Social interaction occurs when two or more individuals, having mutual focus of attention but not necessarily physically co-present, communicate and interact with one another [27]. Examples include face-to-face verbal conversations, email conversations, and non-verbal (sign language) conversations. Goffman [15] distinguishes between focused and unfocused interactions. Focused interaction occurs when two or more co-present individuals, having mutual focus of attention, interact by establishing face-to-face engagement and direct conversation. Note that face-to-face engagement is often not maintained throughout the entirety of a focused interaction. Unfocused interaction, on the other hand, occurs when individuals, though co-present, do not establish a direct engagement and conversation [15]. We use the term conversational partner for a person who is involved in a focused interaction with the camera wearer. In light of Goffman’s theory [15], it is important to highlight that papers describing current egocentric video-based methods [2, 5, 13] that use the term social interaction are actually addressing focused interaction; social interaction is a much broader term.

Figure 1 shows sample frames from three videos in our
Focused Interaction Dataset\footnote{We plan to release the focused interaction dataset after publication}. These examples highlight both the variability in egocentric video data in terms of viewpoint, location, and illumination, and the fact that conversational partners are not always in the field of view (e.g., focused interaction while walking). Voice activity and camera motion cues will be especially important in such cases. Voice Activity Detection (VAD) is widely researched in audio signal processing and used for audio conferencing, speech encoding, speech recognition, and speaker recognition \cite{17, 26}. VAD methods detect voice activity (primarily speech) from a noisy audio signal \cite{16, 24, 29}. Video content-based camera motion analysis methods make use of template matching \cite{11} and optical flow \cite{6}. Methods derived from optical flow are widely used nowadays for human activity and action recognition from third person perspective \cite{8, 20} (where a fixed and static camera captures third person activities such that the optical flow is strongly associated with their activity) and first person perspective \cite{30} (where camera wearer activities affect the global camera motion).

Existing methods for social interaction detection in egocentric videos typically assume that people are already present in the field-of-view of the camera and focus on detecting sub-categories of social interaction \cite{13}, social groups \cite{5}, and presence or absence of social interaction \cite{2} by utilizing visual data only. Fathi et al. \cite{13} proposed one of the first methods for detecting different types of social interaction in egocentric video and evaluated it on clips from videos captured at a theme park. They used a multi-label hidden conditional random field model to detect discussion, monologue and dialogue based on estimates of faces’ locations and orientations. Alleto et al. \cite{5} applied the concept of F-formation \cite{10} for detecting social groups in egocentric video. They designed a pairwise feature vector that describes spatial relationships between two people present based on distances and orientations. A correlation clustering algorithm was applied to merge people into socially related groups. A structural SVM-based method was then used to learn the weight of each component of the correlation clustering vector depending on the social situation. More closely related to our work, Aghaee et al. \cite{2} proposed a method for detecting social interaction in low frame rate photo streams. They trained an LSTM-RNN classifier to detect social interaction based on estimates of the distance of an individual from the camera wearer as well as their relative orientation. These social interaction detection methods \cite{5, 2, 13} processed data offline and considered clips of photo streams that always contained people. However, people interacting with the camera wearer may not always be in the field-of-view (e.g., when walking while having a conversation). Moreover, it should be noted that the existing focused interaction methods only consider constrained video segments (clips), where each clip belongs to one specific class, in which the camera wearer is stationary, hence video cues such as face tracking and orientation alone are sufficient \cite{2}. However, a continuously recorded life-logging video has multiple transitions from one class to another.

In this paper, we address the task of identifying temporal segments in continuous egocentric video that correspond to periods of no-focused interaction (no-FI), focused interaction while the camera wearer is stationary (FI-NW) and focused interaction while the camera wearer is walking (FI-W). All such instances of focused interaction should be automatically detected. Often, FI-W is enclosed within FI-NW (e.g., the camera wearer meets a conversational partner, they go for a walk together, and conclude their interaction with a farewell while facing each other). We propose and evaluate a method based on audio and video features to perform the tasks of detection and classification of focused interaction. The main contributions of this paper are as follows.

- We formulate the task of automatic, online classification of focused interactions in continuous, egocentric audio-video data.
- We use spatio-temporal local and global video features and voice-based audio features for classifying focused interaction.
- We propose a temporal segmentation approach based on frame classification and present several variants of it that use Support Vector Machines (SVM) with either
linear or non-linear kernels for classification using various audio-visual feature sets.

- We evaluate the proposed methods on our Focused Interaction dataset, providing empirical evidence that fusion of visual face track scores, camera motion features and voice activity detector scores, and learning using SVMs with non-linear kernels, provides an effective means for classifying focused interactions.

Note that existing face detection and tracking, voice activity detection and global camera motion analysis techniques are adopted in our proposed method as the aim of this work is not to improve these individual techniques but to look into the effect of integrating these techniques to form a robust and online focused interaction system for unconstrained, life-logging, egocentric videos.

The remainder of the paper is organised as follows. The proposed method for focused interaction classification is detailed in Section 2. Section 3 describes our focused interaction dataset and the evaluation protocol. Results of experiments comparing variants of the methods are presented and analysed in Section 4. Finally, Section 5 draws some conclusions.

2. The proposed method

We process audio and video streams independently to extract three distinct features, namely, face track score, camera motion feature vector and voice activity detection score. From the video stream, the face track score is obtained by detecting and tracking faces, and the camera motion feature vector is obtained by computing the histogram of oriented optical flow. From the audio stream, the voice activity detection score is computed by analysing discriminative audio features. These features are fused to form the feature set and SVMs are then trained for the online classification of continuous data streams that contain instances of No-Fi, FI-NW and FI-W.

2.1. Face track score

A Histogram of Oriented Gradient (HOG)-based face detector [11] is applied to detect faces in each video frame. Some false and missed face detections are inevitable due to the relatively unconstrained nature of egocentric video. Therefore, we apply a Kanade-Lucas-Tomasi (KLT) point tracker [22, 28] to refine face detection results. Tracking is initialized as soon as a face is detected and continues tracking points of the face in subsequent frames. Track points are updated by taking input from the face detector every tenth frame. The face track is terminated if no face is detected at the same position as that of the tracker or if all points that were tracked are eventually lost.

The KLT tracker returns confidence scores for the point tracks. The neighborhood of the \( t^{th} \) point at frame \( t \) consists of those pixels in an image patch \( I^t \) centred on that point. We compute a face tracker score \( s_t \) by summing scores of all points tracked on a face, i.e.,

\[
s_t = \sum_{l=1}^{L} (1 - \frac{1}{W} ||I^t_l - I^t_{l(t-1)}||^2)
\]

where \( W \) is the number of pixels in a neighbourhood patch. The lower bound for \( s_t \) is zero when no faces are tracked while the upper bound depends on the number of points tracked per face (and is certainly no larger than the number of pixels in the face detection box). The track score is high if lots of face points are tracked with confidence. We compute the duration (life) of each track and, in frames in which multiple faces are tracked, we select the one with the longest duration for inclusion in the current feature set. The rationale for this is that short duration tracks often correspond to false detections or to short unfocused interactions (e.g., walking past another person). Moreover, selecting the track with longest duration allows online pruning of tracklets generated through false face detection as these tracklets have comparatively shorter life.

2.2. Camera motion feature extraction

In the case of first person perspective, the distribution of optical flow produces distinct profiles when the camera is static (e.g., camera wearer standing/sitting) and moving (e.g., camera wearer walking, turning around, going up/downstairs). Histogram of Oriented Optical Flow (HOOF) [8] gives a representation of camera motion at each frame.

Given an input video, we compute dense optical flow using Farneback’s method [12]. This results in the flow vectors \( v = [x, y]^T \) and their orientations \( \theta = \arctan \left( \frac{y}{x} \right) \).

HOOF features [8] are computed by binning each flow vector based on its angle with the horizontal axis and weighting it based on its magnitude. The range for the \( b^{th} \) bin is defined as

\[
2\pi \frac{b - 1}{B} \leq \theta < 2\pi \frac{b}{B},
\]

where \( 1 \leq b < B \) and \( B \) is the total number of bins. As a result, we get the normalised histogram \( \hat{h}_t = [\hat{h}_{t,1}, \hat{h}_{t,2}, ..., \hat{h}_{t,B}]^T \) of HOOF features at each time instant \( t \). \( \hat{h}_t \) is then fused with other features in order to form the feature set.

Figure 2(b) shows HOOF features for a sample video along with the activity description. A focused interaction...
occurred between (i) and (iii), and another started at (vii). These are both examples of interactions in the FI-NW class. Note that although FI-NW occurs when the camera wearer is stationary, in an active conversation body motion is present; hence variability in HOOF can be observed. Before (i), the camera wearer was engaged in checking emails on a computer; hence the pattern of HOOF features remained consistent. Between (iii) and (vii), the camera wearer searched for some documents and started walking.

2.3. Audio-based feature extraction

We utilize the method and implementation of Segbroeck et al. [29] for voice activity detection (VAD). This method combines four types of discriminative audio features in order to detect voice activity in noisy real-world environments, specifically, spectral shape, spectro-temporal modulations, harmonicity (presence of pitch harmonics) and long-term spectral variability. The resulting VAD
scores range from 0 to 1; a score close to 0 indicates no
voice activity while a score close to 1 indicates with high
confidence the occurrence of voice activity.

Figure [2](c) shows the estimated VAD scores for a sam-
ple video. At (i) a focused interaction begins; although there
is no face present in the field-of-view of the camera at this
point, voice activity is detected as can be observed in (c).
Another focused interaction begins at (vii) but due to mo-
tion blur and distance of the participant from the camera,
no face is detected until (viii). However, voice activity is
detected from the start of this focused interaction. At other
times ((iv), (v), (vi)), voice activity is falsely detected albeit
with relatively low scores.

2.4. Audio-visual feature fusion

Video and audio features are obtained at different sam-
pling rates; visual features are updated once every video
frame, i.e. at $25\text{Hz}$, whereas VAD scores are computed
every $10\text{ms}$, i.e. at $100\text{Hz}$, given an input audio stream
with sampling rate of $8000\text{Hz}$ (the default setting proposed
by [29]). In order to fuse these features we resample the au-
tdio features, specifically we average four consecutive VAD
scores, with a step size of four, to get the score at the same
rate as that of the video features.

The tracker score $s_t$, HOOF $h_t$ and VAD score $v_t$ are
normalised to have zero-mean and unit variance based on
estimates of their mean and variance obtained from training
data, resulting in $\hat{s}_t$, $\hat{h}_t$ and $\hat{v}_t$. In each frame, the three ex-
tracted features are concatenated to form the feature vector
$f_t = [\hat{s}_t, \hat{h}_t, \hat{v}_t]^T$.

2.5. Focused interaction classification

The task to be performed is to sequentially process the
input audio-visual data stream in order to identify temporal
segments corresponding to periods of No-FI, FI-NW and
FI-W. One way to formulate a solution is to classify each
frame as belonging to one of the three classes. We train
Support Vector Machines (SVM) for classification by using
the features extracted from a fixed-length temporal window
and the ground-truth label for each window.

We trained SVMs on feature vectors which were the con-
catenation of the audio and video features extracted from
each of $M$ consecutive frames. The goal was to assign to
each temporal window of $M$ frames, the classification label
for the frame at the middle of the window. Windows were
extracted with a shift of $H$ frames so that a classification
was obtained every $H^{th}$ frame. LIBSVM [7] was used for
training the SVM using either linear or RBF kernel.

3. Experimental Setup

3.1. Dataset

The number of annotated datasets publicly available for
research that capture social interactions using ego-centric
cameras is limited, in part due to privacy concerns. Some
datasets [21] are only partially available, without audio and
with anonymized (blurred) faces of people in the field-
of-view of the camera. Others captured photostreams at
2 frames per minute of social interaction (without audio) [2]
(not yet publicly available) or of multiple people interacting
as social groups at 4 different locations [5]. Another
dataset captured in a theme park is labelled for three dif-
ferent types of social interaction (dialogue, discussion and
monologue) as well as for activities (e.g., walking, waiting,
gathering, sitting). However, our everyday scenarios are
significantly different from activities performed in a theme
park. Therefore, we recorded a Focused Interaction dataset
that captured various focused and unfocused interactions in-
terspersed naturally with periods of no interaction, in real-
world unconstrained scenarios and in varying environmen-
tal conditions (e.g. indoor/indoor, daylight/night).

Our Focused Interaction Dataset contains 19 egocentric
continuous videos captured, at high resolution (1080p) and
at a frame rate of 25fps, using a shoulder-mounted GoPro
Hero4 camera and a smartphone (for inertial and GPS data),
at 18 different locations and with 16 different conversational
partners. This makes our dataset useful for other egocentric
applications such as scene categorization and person associ-
ation. A shoulder-mounted camera is preferred over a head-
mounted one as it is less obstructive and provides relatively
stable video because the camera does not move with the
user’s head. Our dataset contains 378 minutes of recordings
(approximately 560k video frames) annotated into periods
of No-FI, FI-NW and FI-W. The dataset is unconstrained
in nature as neither the camera wearer nor the conversa-
tional partners were given any specific instructions that may
restrict their movement and it was captured at several in-
door and outdoor locations at different times of the day and
night, and in different environmental conditions (e.g. sunny
or cloudy, with background noise from nearby people and
cars). In total, 240 mins (64%) of data contain focused in-
teraction in which conversational partners are in the field-
of-view of the camera most of the time and are not walk-
ing; their positions and face orientations vary significantly.
50 mins (13%) of data contain focused interactions in which
the conversational partners are not in the field-of-view of
the camera and 88 mins (23%) of data do not contain any
focused interaction.

3.2. Evaluation protocol

We used seven different feature sets for evaluation, namely,
TVM, TV, TM, VM, T, V and M, where T de-
Table 1: Evaluation of various feature sets and SVM kernels when using one-versus-all SVM classification. Values given are pooled over the 6 validation folds. Key: $P$ - precision; $R$ - recall; $F$ - F1-score; AUC - area under curve; $T$ - track score; $V$ - VAD score; $M$ - motion vector; FI - focused interaction; NW - non-walk; W - walk.

<table>
<thead>
<tr>
<th>Feature set</th>
<th>Class</th>
<th>Linear kernel</th>
<th>Non-linear (RBF) kernel</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$C$</td>
<td>$P$</td>
</tr>
<tr>
<td>TVM</td>
<td>No-FI</td>
<td>2^{-9}</td>
<td>93.54</td>
</tr>
<tr>
<td></td>
<td>FI-NW</td>
<td>85.16</td>
<td>91.59</td>
</tr>
<tr>
<td></td>
<td>FI-W</td>
<td>91.30</td>
<td>97.04</td>
</tr>
<tr>
<td>TV</td>
<td>No-FI</td>
<td>2^{-6}</td>
<td>93.65</td>
</tr>
<tr>
<td></td>
<td>FI-NW</td>
<td>83.79</td>
<td>93.39</td>
</tr>
<tr>
<td></td>
<td>FI-W</td>
<td>89.59</td>
<td>98.02</td>
</tr>
<tr>
<td>TM</td>
<td>No-FI</td>
<td>2^{-9}</td>
<td>83.23</td>
</tr>
<tr>
<td></td>
<td>FI-NW</td>
<td>83.28</td>
<td>92.90</td>
</tr>
<tr>
<td></td>
<td>FI-W</td>
<td>86.94</td>
<td>98.30</td>
</tr>
<tr>
<td>VM</td>
<td>No-FI</td>
<td>2^{-9}</td>
<td>90.82</td>
</tr>
<tr>
<td></td>
<td>FI-NW</td>
<td>69.79</td>
<td>88.51</td>
</tr>
<tr>
<td></td>
<td>FI-W</td>
<td>86.81</td>
<td>99.99</td>
</tr>
<tr>
<td>T</td>
<td>No-FI</td>
<td>2^{0}</td>
<td>79.19</td>
</tr>
<tr>
<td></td>
<td>FI-NW</td>
<td>82.33</td>
<td>93.11</td>
</tr>
<tr>
<td></td>
<td>FI-W</td>
<td>86.81</td>
<td>99.90</td>
</tr>
<tr>
<td>V</td>
<td>No-FI</td>
<td>2^{-10}</td>
<td>91.08</td>
</tr>
<tr>
<td></td>
<td>FI-NW</td>
<td>67.53</td>
<td>59.65</td>
</tr>
<tr>
<td></td>
<td>FI-W</td>
<td>86.81</td>
<td>100.00</td>
</tr>
<tr>
<td>M</td>
<td>No-FI</td>
<td>2^{-9}</td>
<td>76.82</td>
</tr>
<tr>
<td></td>
<td>FI-NW</td>
<td>62.20</td>
<td>28.99</td>
</tr>
<tr>
<td></td>
<td>FI-W</td>
<td>86.81</td>
<td>100.00</td>
</tr>
</tbody>
</table>

notes the face track score, $V$ denotes the VAD score and $M$ denotes the camera motion feature vector. SVM with either linear or RBF kernel in one-versus-all setting was used for the classification using different feature sets. For each feature set, the best performing SVM parameters, $C$ and $\gamma$, for linear and RBF kernels were computed using grid search [18] (reported in Table 1) and were then used for the validation. A temporal window size of 50 frames (selected empirically) with a shift of 25 frames was used for obtaining the training samples, while testing was performed with a shift of 1 frame.

Evaluation was performed using six-fold division of our focused interaction dataset. Since the duration of each recording varied, it was not possible to have exactly equal numbers of frames in each fold without arbitrarily breaking videos into smaller parts. Instead the folds were generated to roughly contain 60 mins of data.

We use framewise evaluation measures to assess the performance by comparing the predicted labels against the ground-truth labels. For each class, we plot the Receiver Operating Characteristic (ROC) curve using the one-versus-all strategy and compute the Area Under the Curve (AUC). The Precision, $P$, Recall, $R$, and F1-score, $F$, are then reported. For the three-class confusion matrix, the predicted class labels are obtained by assigning a positive label to the class with maximum score among the three classes.

4. Results and Discussion

Table 1 summarises the one-versus-all evaluation results for the different feature sets using either linear or RBF kernel for SVM. The corresponding ROC curves are shown in Fig. 3.

TVM-RBF outperformed other feature sets and linear kernel giving an AUC of 94.65% for No-Fi, 96.46% for FI-NW and 94.53% for FI-W, respectively. TVM-RBF gave F1-score of 94.46% for No-Fi, 89.01% (slightly lower than F1-score for TV-RBF) for FI-NW and 95.62% for FI-W, respectively. In the case of No-Fi, voice activity and face tracks are not present most of the time but different types of camera motion occur, e.g., camera wearer walking, sitting or standing alone. Hence motion feature (M) alone cannot reliably identify this class. Observe the low performance of M in Fig. 3(a), (d). From these figures and Table 1 it can be observed that the performance is comparable when using TVM-Linear and TV-Linear for No-Fi, while the use
of RBF kernel helped to improve the performance (AUC increased by 2% for TVM). The instances of FI-NW contain face-to-face interaction, where face tracks are present most of the time. Hence TVM, TV and T with both linear and RBF kernels performed equally well (Fig. 3(b), (e)). M provides a strong cue for discriminating between FI-NW and FI-W. Face tracks are not present in FI-W as the conversational partner is not in the field-of-view of the camera. Moreover, FI-W occurred mostly in outdoor scenarios where the audio signal might get corrupted by background noise (from roadside and passing-by people). Hence, the performance of FI-W is extremely low when using T or V alone. It can be observed from Fig. 3 (comparing the 1st row with the 2nd row) that M has a non-linear relation with the focused interaction classes. This is evident from lower performance of M when using a linear kernel but significantly improved performance when using a non-linear kernel.

The confusion matrices for top performing feature sets are shown in Fig. 4. A confusion matrix is computed by selecting the class with maximum score at each frame in the one-versus-all strategy. TVM-Linear gave accuracy of 62.3% for No-FI, 95.1% for FI-NW and 70.6% for FI-W (Fig. 4(a)). TVM-RBF outperformed by giving accuracy of 73.3% for No-FI, 93.3% for FI-NW and 80.7% for FI-W (Fig. 4(f)). RBF kernel with M, in particular, is useful for classifying FI-W as this interaction occurred while walking. As observed from the ROC curves (Fig. 3), TVM and TV gave comparable results for FI-NW as motion features do not contribute much. It can be observed from the confusion matrices that the overall accuracy when using TV as feature set with linear kernel was 80.9% (Fig. 4(b)) and with RBF kernel was 85.5% (Fig. 4(g)). Similar trend is observed when using TVM as feature set giving an overall accuracy of 82.6% (Fig. 4(a)) with linear kernel and 86.8% (Fig. 4(f)) with RBF kernel, suggesting that the use of non-linear SVM kernel helped in improving the performance.

5. Conclusion

We have presented a method for the automatic online classification of focused interaction in continuous, egocentric videos captured in unconstrained everyday scenarios.
Figure 4: Confusion matrices obtained using different feature sets and SVMs with linear (upper row) and non-linear (lower row) kernels.

We processed both audio and video data streams to obtain audio-visual feature sets. In particular, fusion of face track scores and camera motion profile extracted from visual data with voice activity detection scores from audio data proved to be effective. We performed temporal segmentation of focused interactions via classification using SVMs with different kernels. We evaluated variants of the methods, including single and multimodal feature sets. The use of camera motion profile along with face track and voice activity detection scores and SVM with non-linear kernel were in particular useful for discriminating no focused interaction and focused interaction while walking. Face track and voice activity detection scores were significant for discriminating face-to-face focused interaction for which SVM with non-linear kernel and camera motion profile did not give any further improvement.

In future, we plan to extend this work to identify conversational partners even when they are not in the field-of-view of the camera (e.g., focused interaction while walking) to enhance assistive technology for non-speaking people wearing an egocentric camera.

Acknowledgements. This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) under grant EP/N014278/1: ACE-LP: Augmenting Communication using Environmental Data to drive Language Prediction. The authors are grateful to Annalu Waller (University of Dundee), the ACE-LP team and CVIP members (University of Dundee) for useful discussions and assistance with dataset collection.

References


