IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2018.DOI

Multimodal Egocentric Analysis of Focused Interactions

SOPHIA BANO, TAMAS SUVEGES, JIANGUO ZHANG, (Senior Member, IEEE), AND STEPHEN J. MCKENNA, (Senior Member, IEEE)

CVIP, School of Science and Engineering, Queen Mother Building, University of Dundee, Dundee, DD1 4HN, Scotland, UK
Emails: {s.bano, t.suveges, j.n.zhang, s.j.z.mckenna}@dundee.ac.uk

Corresponding author: Stephen J. McKenna (e-mail: s.j.z.mckenna@dundee.ac.uk).

This work was 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*.

ABSTRACT Continuous detection of social interactions from wearable sensor data streams has a range of potential applications in domains including health and social care, security, and assistive technology. We contribute an annotated, multimodal dataset capturing such interactions using video, audio, GPS and inertial sensing. We present methods for automatic detection and temporal segmentation of focused interactions using support vector machines and recurrent neural networks with features extracted from both audio and video streams. Focused interaction occurs when co-present individuals, having mutual focus of attention, interact by first establishing face-to-face engagement and direct conversation. We describe an evaluation protocol including framewise, extended framewise and event-based measures and provide empirical evidence that fusion of visual face track scores with audio voice activity scores provides an effective combination. The methods, contributed dataset and protocol together provide a benchmark for future research on this problem. The dataset is available at https://doi.org/10.15132/10000134.

INDEX TERMS Social interaction, egocentric sensing, multimodal analysis, temporal segmentation.

I. INTRODUCTION

We consider automatic detection of social interactions by analysis of wearable sensor data. Specifically, we address the problem of identifying periods during which the wearer of the sensors, the subject, is involved in focused interaction. To support work in this area, we provide an annotated, multimodal dataset, an evaluation protocol, and results from methods that sequentially parse audio-visual streams to serve as a baseline for future research.

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 [1]. Face-to-face engagement is often not maintained throughout the entirety of a focused interaction; for example a group of people talking while in conversation will typically look at each other only intermittently. This concept of focused interaction is more specific than that of *social interaction* which can be considered to occur whenever individuals communicate and interact with one another whether or not they are physically co-present, e.g. by telephone [2]. In particular, the category focused interaction excludes unfocused interactions in which individuals, though co-present, do not establish a direct engagement and conversation [1]. Individuals in an unfocused interaction are aware of each others' presence but establish only indirect engagement which might involve brief eye contact, or facial expressions for example.

Automatic identification of a subject's focused interactions has various potential applications such as in behaviour understanding for health and social care [3], [4], evidence management for security and law enforcement (video 'badges') [5]–[7], and as a precursor to more fine-grained analysis of interactions. In order to facilitate and encourage research on this problem we provide a multimodal dataset that includes one-to-one interactions as well as group interactions, in a range of indoor and outdoor scenarios. All focused interactions and unfocused interactions are annotated. We present results on this dataset for one particular task, that of sequential recognition of focused interactions based on audio and visual data streams. Figure 1 shows video frames from four of the focused interactions in the dataset. These examples highlight the variability of viewpoint, location, and illumination, and the fact that interaction partners are not always in the field of view. Audio cues will be especially important in such cases.

This paper extends our preliminary system for discriminat-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2850284, IEEE Access

(a) in view, outdoors

(b) in view, indoors

(c) not in view, outdoors

(d) not in view, outdoors at night

FIGURE 1: Video frames from four focused interactions.

ing interactions while walking and not walking using audio-visual features [8]. We report results for detecting focused interactions using more data, temporal filtering, and Long Short-Term Memory (LSTM) recurrent neural networks as well as Support Vector Machines (SVMs) using audio-only, video-only, and audio-visual features. Furthermore, we compare directly with an implementation of a related system outlined by Hayden [9]. It is not an aim of this paper to propose novel algorithms for sub-tasks such as face detection, face tracking and voice activity detection per se; rather we investigate their integration to form an effective system. Additionally, we characterise performance in detail using frame-wise, extended frame-wise and event-based measures. The proposed methods, the contributed dataset and the evaluation protocol together provide a benchmark for detection and analysis of focused interactions in ego-centric data. The dataset should also be useful for investigating tasks such as location and person association.

II. RELATED WORK

Audio is an integral part of social interaction as voice activity is prevalent during such interaction. Voice Activity Detection (VAD) is widely researched in audio signal processing and useful for several applications such as audio conferencing, speech encoding, speech recognition, and speaker recognition [10], [11]. VAD methods detect voice activity (primarily speech) from a noisy audio signal [12]–[14]. In real-world videos, VAD is a challenging task as the associated audio signals are usually degraded due to noise from the surroundings.

Social interaction detection has been investigated using computer vision. Much of this is from a third-person perspective [15]–[17] but there is also work on detection and analysis of interaction from a first-person perspective using wearable cameras. Egocentric video is relatively unconstrained in nature as it is recorded from a non-static camera worn on the head or body of a person [18]. In contrast with most third-person perspective video in which the focus of attention is usually well captured within the camera's field of view, in first-person perspective video the focus of attention may not always lie in the field of view and the viewpoint varies a lot. Moreover, life-logging style video is captured in varied environments in both indoor and outdoor locations (e.g. parks, restaurants, offices, cars, tourist attractions), at day or night, and in varied weather conditions. These characteristics often make automatic analysis of egocentric video more challenging. Methods developed for third-person perspective video are often not directly applicable. Analysis of egocentric sensor data has gained the attention of researchers for tasks such as object recognition [19], [20], activity recognition [21], temporal segmentation [22], [23], video summarisation for life-logging [18], person-to-person (person-to-group) interaction recognition [24]–[26] and person-to-object interaction recognition [21], [27]. 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 [28], [29], social interaction detection in nursing homes using surveillance-type camera videos [30] and scene change detection in life-logging videos [31].

Methods have been proposed to detect groups of individuals interacting with each other or with the camera wearer [24], [26]. Fathi *et al.* [24] presented the first study detecting different types of social interaction in egocentric video and performed evaluation on data captured at a theme park (see Table 1). They used a multi-label hidden conditional random field model to detect discussion, monologue and dialogue based on estimates of faces' locations and orientations. Building on earlier work that used the concept of F-formation in the analysis of third-person perspective videos captured from static cameras [16], [32], Alleto *et al.* [25] applied that concept to detecting social groups in ego-centric video using the Ego-Group dataset (Table 1). They designed a pairwise feature vector that describes spatial relationships between two people based on distances and orientations. A correlation clustering algorithm was used to merge people into socially related groups and a structural SVM-based method was used to learn the weight of each component of the clustering vector depending on the social situation. Aghaei *et al.* [26] proposed a method for detecting social interaction in low frame-rate photo streams (UB social interaction dataset - not publicly available). They trained an LSTM recurrent neural network to detect social interaction based on estimates of the distance of an individual from the camera wearer and their relative orientation. They further extended this work [26] to social style characterisation [33], in which distance, orientation and facial emotion [34] were used for social interaction detection; facial emotion and environmental (dimensionality-reduced VGG-NET) features were used to classify an interaction as formal or informal. These existing social interaction detection methods [24], [26], [33] processed data offline and considered short video clips or photo streams captured from constrained perspectives that always contained people. In this paper we process long, continuous sequences in which conversational partners are not always in the field-of-view during interaction.

SVM classifiers are often used for human activity recognition based on spatio-temporal features [35], [36]. Recurrent Neural Networks (RNNs) can represent and make use of arbitrarily lengthy historical data and are able to exhibit dynamic temporal behaviour. They have also been used with some success for human activity recognition [26], [33], [37]–[39]. For example, Hammerla et al. [37] used RNNs to recognise activities from wearable device data, Abebe and Cavallaro [38] used LSTM-RNN for egocentric ambulatory human activities recognition from pre-segmented video clips of individual activities, and DeepSense [39] integrated Convolutional Neural Networks (CNN) and RNN for solving both the regression and classification oriented online mobile sensing problems. By analysing the performance of both SVM and LSTM-RNN with audio, visual, and audio-visual features, we aim to obtain a deeper understanding of our application and dataset, and to provide more comprehensive benchmarking for future research.

III. FOCUSED INTERACTION DATASET

A. RELATED DATASETS

The number of annotated datasets publicly available for research that capture social interactions using first-person cameras and other body-worn sensors is limited. Several datasets have been captured for related purposes such as engagement detection [40] and object interaction [20]. Here we report those datasets acquired with similar aims in mind to ours (summarised in Table 1). The UT Ego dataset contains recording of daily activities which include interactions with friends [19]. However, only 4 of 10 videos without audio have been made available with people anonymised by blurring their faces for privacy reasons. The UB Social Interaction dataset contains photo streams without audio captured at 2 frames per minute using a narrative camera [26]. Their duration varied from 5 to 20 minutes (10 to 40 frames) and each stream always contained at least one individual either interacting or not interacting with the camera wearer. The Ego Group dataset contains multiple short photo streams without audio totaling 2900 frames (116 *secs*) that capture multiple people interacting as social groups in different situations: in a laboratory, at coffee break, in a conference room and in an outdoor setting [17]. The First-Person Social Interaction dataset contains day-long videos of multiple peoples' experience of visiting a theme park [24]. These videos are labelled for three different types of social interaction (dialogue, discussion and monologue) and for activities such as walking, waiting, gathering, sitting, buying something, and eating. However, these activities and interactions occurred in a relatively unusual setting; our everyday scenarios are significantly different from activities performed in a theme park. Moreover, this dataset has focused or unfocused interaction throughout its entirety as the camera wearer was always accompanied by a partner. In everyday scenarios, we are not necessarily accompanied throughout the day. We meet and interact with certain people often in some particular locations (e.g. breakfast with family, greeting colleagues at workplace) but these interactions do not last all day long.

We contribute a *Focused Interaction* dataset that, unlike existing datasets, continuously captures various interactions interspersed naturally with periods of no interaction, in real-world unconstrained scenarios and in varying environmental conditions (e.g. indoor/outdoor, day/night) using video, audio, inertial sensing, and GPS.

B. SENSORS

We carried out initial feasibility trials with four different wearable camera set-ups: an Edesix VB-300 camera[2] with shirt pocket mount, a head-mounted GoPro Hero 4, a shoulder-mounted GoPro Hero 4, and Vuzix M100 smart glasses[3]. All these camera set-ups captured video from an ego-centric perspective. The M100 and VB-300 often failed to capture the focus of interest due to narrow field of view

[2]Edesix VB-300: https://www.edesix.com/products/vb-300, last accessed: 10062018.

[3]Vuzix M100: https://www.vuzix.com/products/m100-smart-glasses, last accessed: 10062018

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2850284, IEEE Access

TABLE 1: Related egocentric datasets concerned with social interaction. Key: NS - not specified; RES - frame resolution; V_{fps} - video frame rate.

Dataset name	Camera	Mount	Video released	Audio released	Blurred faces	RES	V_{fps}	Annotation	Description
UT Ego Dataset[a] [19]	Looxcie camera	Around ear	4/10 videos	×	✓	320 × 480	15	Objects of interest (OI) or not, binary mask for regions with OI	Continuous videos that capture daily activities such as eating, shopping, attending a lecture, driving, and cooking.
UB Social Interaction Dataset[b] [26]	Narrative camera	Body	✓ (images)	×	×	512 × 385	2*fpm*	Bounding boxes around faces in each frame	Multiple short photo streams always containing one or more individuals at times involved in interaction with the camera wearer.
Ego Group Dataset[c] [17]	NS	NS	✓	×	×	960 × 540	NS	Group number assigned to each person in a frame	Multiple short video clips that always contain people, capturing social groups in four situations: laboratory, coffee break, conference room and outdoor scenario.
First-Person Social Interaction Dataset[d] [24]	GoPro	Head cap	✓	✓	×	1280×720	30	Activity type, social interaction type	Continuous videos captured at a theme park containing activities (walking, waiting, sitting, buying, eating, etc) and social interactions labeled as dialogue, discussion and monologue.
Focused Interaction Dataset[e]	GoPro Hero4	Shoulder	✓	Voice activity	×	1920 × 1080	25	Focused, unfocused and no interaction along with person ID	Continuous videos captured at various locations (indoor, outdoor, day & night time, office, campus). Focused and unfocused interactions with multiple people interspersed with periods of no interaction.

[a]UT Ego Dataset: http://vision.cs.utexas.edu/projects/egocentric_data/UT_Egocentric_Dataset.html (last accessed: 10-06-2018)
[b]UB Social Interaction Dataset: http://www.ub.edu/cvub/dataset/egosocialstyle/ (last accessed: 19-06-2018).
[c]Ego Group Dataset: http://giuseppeserra.com/content/egocentric-vision-detecting-social-relationships (last accessed: 10-06-2018)
[d]First Person Social Interaction Dataset: http://ai.stanford.edu/~alireza/Disney/ (last accessed: 10-06-2018)
[e]Focused Interaction Dataset: https://doi.org/10.15132/10000134

and had unwanted jitter motion when the camera wearer was walking due to semi-rigid mounts and lack of optical stablisation. GoPro, on the other hand, comes with an inbuilt optical stabilisation which compensates for the unwanted camera motions, and has a wider field-of-view. The recorded video quality of GoPro is better than the M100 and VB-300 as it captures sharp videos with high resolution. A head-mount enabled capturing head motion along with the body motion but its appearance was bulky making the camera wearer uncomfortable both in terms of the added load on the head and unnecessary attention from passers-by. A shoulder-mount captured only the body motion but it was less obstructive to the camera wearer and preferred in terms of comfort and ease of use. Due to the wide-angle of the GoPro, even with the shoulder-mount, the subject's focus of interest was captured most of the time. For these reasons we used a shoulder-mounted GoPro Hero 4 to capture audio and video. We also used a smartphone (placed in the camera wearer's right-hand trouser pocket) to capture GPS, accelerometer and gyroscope data. The Androsensor[4] android app was installed on the smartphone logging sensory data to a csv file on the SD card.

C. DATASET COLLECTION PROCEDURE

We obtained ethical approval to record at a range of indoor and outdoor locations on campus. Written consent was obtained from each conversational partner. No children were recorded, either intentionally or unintentionally. The subject wore a badge stating that recording was being undertaken for research purposes. Conversational partners were not given any specific instructions to restrict their movement; they were simply asked to have routine interactions.

At the beginning of each recording session, the AndroSensor app and the camera record buttons were turned on. Since these devices started recording at slightly different times, the data streams from the two devices were not synchronised. A synchronisation pulse was therefore generated with a clap action performed by the subject while holding the smartphone in the hand and in front of the camera. This pulse can be used for alignment if data from both devices are to be used. The smartphone was then placed in the subject's trouser pocket. The subject then walked around, meeting conversational partners at various indoor and outdoor locations. The AndroSensor app and the camera were left on continuously throughout each session.

D. DATASET DETAILS

Table 2 details characteristics of the collected focused interaction dataset. This dataset contains 377 minutes (including 566,000 video frames) of multimodal recording including periods in which the wearer is engaged in focused interactions, unfocused interactions, and no interaction. Experiments in this paper use only video and audio data; data from the other sensors was recorded so that it might be incorporated in future studies. In order to introduce diversity in the dataset, recordings were captured while visiting 18

[4]Androsensor: https://play.google.com/store/apps/details?id=com.fivasim.androsensor (last accessed: 10062018)

TABLE 2: Details of the dataset. Key: CP - conversational partner; FOV - field of view; fps - frames per second.

Description	Value
Total number of sessions	19
Total duration of sessions	377.37 *mins*
Minimum session length	6.37 *mins*
Maximum session length	51.83 *mins*
Focused interaction duration (CP in FOV)	239.77 *mins*
Focused interaction duration (CP not in FOV)	49.68 *mins*
Unfocused and no interaction duration	87.92 *mins*
Number of subjects (camera wearers)	1
Total number of CPs	17
Total number of locations	18
Total number of focused interaction instances	145
Video frame rate	25 *fps*
Video resolution	1920×1080
Audio sampling rate	48 *kHz*

different indoor 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). Videos were recorded at 25 *fps* with 1080p resolution and 48 *KHz* audio sampling rate. In total, 19 separate sessions were recorded. The duration of sessions varied and depended on the will of the subject to record scenarios in which they felt comfortable. The shortest session was 6 *mins* and included one focused interaction. The longest session was 52 *mins* and included 16 focused interactions. In total, there are 240 *mins* of focused interactions in which conversational partners are in the field-of-view most of the time (e.g., Figure 1(a) and (b)); their positions and face orientations vary significantly. There are 50 *mins* of focused interactions in which the conversational partners are not in the field-of-view (e.g. while walking as in Figure 1(c) and (d)). The remaining 88 *mins* contain either unfocused or no interaction. Variations in background and face orientation as exemplified in Figure 1 present a challenge to face detectors and trackers.

E. DATASET ANNOTATION

Annotation was performed by the subject; having been involved directly in all interactions we believe the subject is the best person to judge when interactions begin and end. A sample video was also annotated by an independent observer for calibration purposes. The subject used the ELAN tool[5] to label all focused and unfocused interactions. This entailed marking the points in time at which a person joins or leaves an interaction. Additionally, transitions between stationary interactions and interactions while walking were marked. Anonymised IDs for people involved in each focused interaction were also provided.

IV. METHODS

We perform sequential processing of both audio and video streams simultaneously to obtain audio-visual feature vectors (see Figure 2) upon which our models are trained for online detection of focused interaction in continuous data streams.

A. VISUAL FEATURE EXTRACTION

We use a Histogram of Oriented Gradient (HOG)-based face detector in each frame [41]. Given the relatively unconstrained nature of egocentric video, some false face detections and missed detections are inevitable. Therefore, we use Kanade-Lucas-Tomasi (KLT) point tracking to refine face detection results. KLT tracking is more precise than alternatives such as mean-shift face tracking because it tracks multiple corner features which provides a certain robustness against tracking failures [42]. As soon as a face is detected, tracking is initiated to track points on the face in subsequent frames. The points to be tracked are refreshed by getting input from the face detector every tenth frame. If the face detector outputs a face bounding box that overlaps with the tracker bounding box, the points are updated and tracking continues. Alternatively, the 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 lost.

The KLT tracker returns confidence scores for the point tracks. These scores are computed based on the similarity of the neighborhood of a tracked point in the current frame with its neighbourhood in the previous frame. We compute a face tracker score (denoted as T) by summing the confidence scores of all points tracked on a face [8]. In the absence of a track this score is zero whereas in the presence of a track it takes the value obtained by accumulating the confidence scores. It 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. Where multiple faces are tracked, only the face with the longest track duration is selected for inclusion in the current feature set as short duration tracks often correspond to false detections or brief unfocused interactions (e.g., walking past another person). Although our approach gave reliable face tracks, it is worth mentioning that a library such as OpenPose[6] can also be useful for solving this problem.

In addition to track score, we experimented with two other visual features extracted from the face track of longest duration. These were the face detection score (F) returned by the frame-wise face detector and the height (H) of the tracked face bounding box. Clearly height carries information about the distance of the tracked face from the subject.

B. VOICE ACTIVITY DETECTION

We utilise the method and implementation of Segbroeck et al. [13] for Voice Activity Detection (VAD). It combines four types of discriminative audio features 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

[5]ELAN: https://tla.mpi.nl/tools/tla-tools/elan/, last accessed: 10062018.

[6]https://github.com/CMU-Perceptual-Computing-Lab/openpose

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2850284, IEEE Access

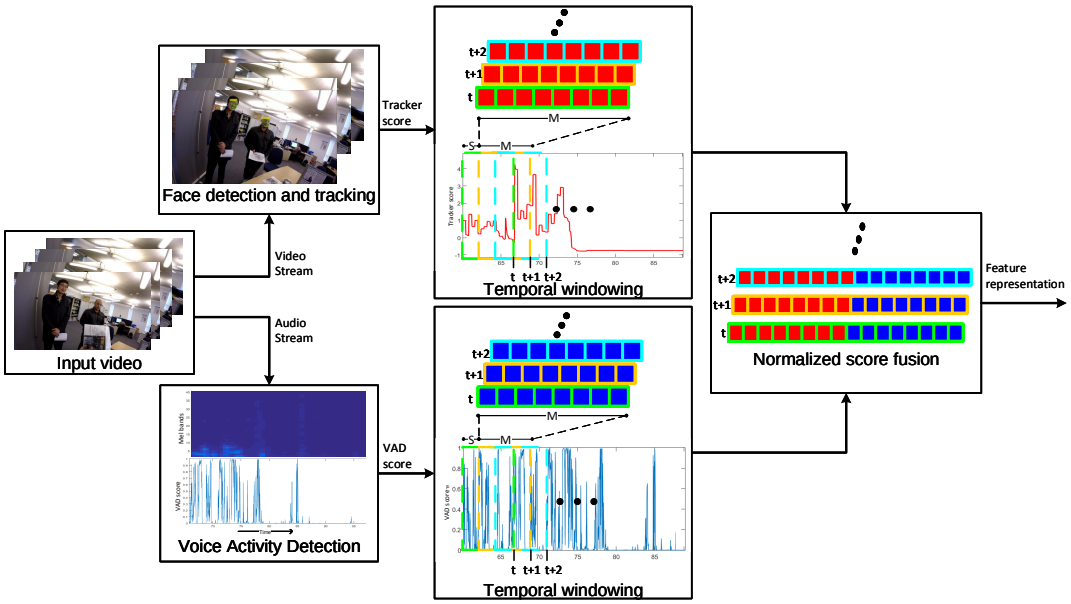


FIGURE 2: Overview of audio and visual feature extraction, temporal windowing and feature fusion. (Best viewed in colour.)

resulting VAD scores (denoted as V) range from 0 to 1; a score at 0 indicates no voice activity while a score close to 1 indicates with high confidence the occurrence of voice activity.

C. AUDIO-VISUAL FEATURE FUSION

Visual and audio features are obtained at different sampling rates; visual features are updated once every video frame, i.e. at 25 Hz, whereas VAD scores are computed every 10 ms, i.e. at 100 Hz, given an input audio stream with sampling rate of 8 kHz (the default setting proposed by [13]). The captured audio stream is down-sampled from 48 kHz to 8 kHz for input to the VAD algorithm. In order to fuse these features we resample the audio 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.

In order to fuse the different audio and video features, each feature is normalised to have zero-mean and unit variance based on estimates of its mean and variance obtained from training data. The features are then concatenated to form a feature vector for each frame.

D. TEMPORAL SEGMENTATION OF FOCUSED INTERACTION

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 focused interaction. One way to formulate a solution is to classify each frame as either belonging or not belonging to a focused interaction. We tried two methods to achieve this binary classification of frames: (i) classification using a Support Vector Machine (SVM) based on features extracted from a fixed-length temporal window, and (ii) classification using a recurrent neural network with Long Short-Term Memory (LSTM-RNN) based on relevant information remembered from features extracted from frames up until the current frame.

1) Sliding window SVM classification

We train linear SVMs on feature vectors which are the concatenation of the audio and video features extracted from each of M consecutive frames. The goal is to assign to each such temporal window the class label of the last frame of that window. We choose to predict the label for the last frame (rather than the middle frame) in order to obtain a low latency method; this is however more challenging and likely to increase fragmentation errors. Windows are extracted with a stride of S frames resulting in a classification at every H^{th} frame (Figure 2). We used the Matlab implementation of L2-regularised SVM (calibrated) with dual solver.

2) LSTM-RNN classification

In order to train an LSTM-RNN, we construct batches of size $k = 4$ as in [37]. A training set consists of multiple videos of various durations. We begin by selecting k training videos at random. Audio-visual feature vectors are extracted from temporal windows of length M at the beginning of these videos and form the first training batch, b_1. Subsequent batches are formed by moving the temporal windows forward in time by M frames for each new batch. Whenever the end of a video is reached, it is replaced by another video selected at random from those not yet used for training in the current epoch. This batch formation process continues until batches have been formed across all the training data; training on all these batches constitutes an epoch.

Whenever the end of a video is reached and replaced by another video from which windows begin to be processed,

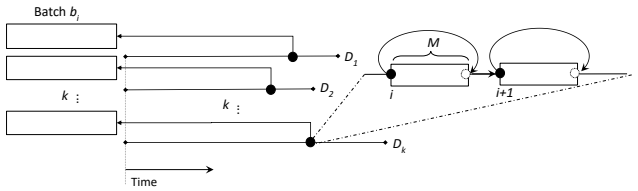


FIGURE 3: Batch formation for LSTM-RNN training from multiple videos. Feature vectors are extracted from temporal windows of length M. The windows move by M frames at each new batch to form b_i. D_k denotes the k^{th} training video.

the state of the LSTM is reset; this avoids learning across the discontinuous transition from one video to another. Training an RNN on a long sequence can result in it essentially memorizing the whole sequence. The method described above avoids this by resetting the state of the RNN at the end of a video. At the start of RNN training we used a learning rate of 0.1 and decreased this by a factor of 10 in each epoch. The network typically reached convergence after 3 epochs. The batch formation process is illustrated in Figure 3.

3) Temporal filtering

Directly thresholding classifier output at a predefined threshold can result in classification errors due to the short-term fluctuations in the output. We experiment with applying a temporal median filter (with empirically selected window size of 145 frames) to smooth such fluctuations.

E. BASELINE METHOD

We implemented the method of [9] as faithfully as the level of detail provided in that thesis allowed. Note that the dataset used in [9] has not been made available. This method is similar in that it uses both audio and video features and is motivated by continuous life-logging in real-world scenarios. In contrast, methods such as [25], [26] used only visual features, were designed for evaluation on short photo streams or video clips always containing people, and rely on face tracks being present during interaction [26] or emphasise analysis of groups of several people interacting [25].

Face presence, size, and head pose features were obtained using the method of Zhu and Ramanan [43]. Head pose estimation returned head orientation quantised to 15 degree intervals in the range -90 degrees to +90 degrees. The audio stream was divided into intervals such that each interval contained audio samples equivalent to one video frame. For each interval, basic energy statistics [44] were computed such as mean, standard deviation, average absolute difference (between the sample values in each interval) and the 10-binned distribution of the sample values in each interval. For the 10-binned distribution, the range (maximum-minimum) of the whole audio stream was first computed and then divided into equal bins. These visual and audio features were normalised and temporal segmentation of focused interaction was performed as detailed in Sec. IV-D using an SVM with a linear kernel.

V. EVALUATION PROTOCOL

We perform 6-fold cross-validation to estimate expected performance. Since sessions are of varying duration it is not possible to have exactly equal numbers of frames in each fold without breaking up sessions arbitrarily into smaller parts. Instead sessions are assigned to folds in such a way as to obtain folds of approximately equal total size ($c.$ 60 *mins*). We use standard framewise, extended framewise, and event measures.

A. STANDARD FRAMEWISE MEASURES

At a chosen operating point (obtained by thresholding the output at 0.5 for both SVM and LSTM), precision - $\mathcal{P}$, recall - $\mathcal{R}$ (true positive rate), F1-score - $\mathcal{F}$ and fall-out - $\mathcal{O}$ (false positive rate) are computed from the confusion matrix. We plot the Receiver Operating Characteristic (ROC) curve and compute the Area Under the Curve (AUC). We also report the Equal Error Rate (EER).

B. EXTENDED FRAMEWISE MEASURES

Extended framewise measures are computed by first dividing the ground-truth and predicted label streams into segments such that a new segment is marked whenever a change occurs in either stream. False positive segments are categorised as insertion errors, merge errors (joining two true positive segments), overfill at start errors (a detection starts too early), and overfill at end errors (a detection ends too late). Extended framewise measures corresponding to these categories can then be defined as the proportion of negative frames in each category [45]:

$$ir = \frac{I_f}{N}, \quad mr = \frac{M_f}{N}, \quad o^\alpha = \frac{O_f^\alpha}{N}, \quad o^\omega = \frac{O_f^\omega}{N} \tag{1}$$

where I_f, M_f, O_f^α and O_f^ω are the numbers of frames in the insertion, merge, overfill at start, and overfill at end categories, and N is the total number of negative labels, i.e., $N = I_f + M_f + O_f^\alpha + O_f^\omega + TN$. Similarly, false negative errors are categorised as deletion errors, fragmentation errors (between two true positive segments), underfill at start errors (a detection starts too late), and underfill at end errors (a detection ends too early). The corresponding extended framewise measures for false negatives are:

$$dr = \frac{D_f}{P}, \quad fr = \frac{F_f}{P}, \quad u^\alpha = \frac{U_f^\alpha}{P}, \quad u^\omega = \frac{U_f^\omega}{P} \tag{2}$$

where D_f, F_f, U_f^α and U_f^ω are the number of frames in the *deletion*, *fragmentation*, *underfill at start* and *underfill at end* categories, and P is the total number of positive labels, i.e., $P = D_f + F_f + U_f^\alpha + U_f^\omega + TP$.

C. EVENT MEASURES

An *event* is a contiguous segment of positive frames, in either the ground-truth labelling or in the predicted labelling. Ground-truth and prediction events can be categorised with

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2850284, IEEE Access

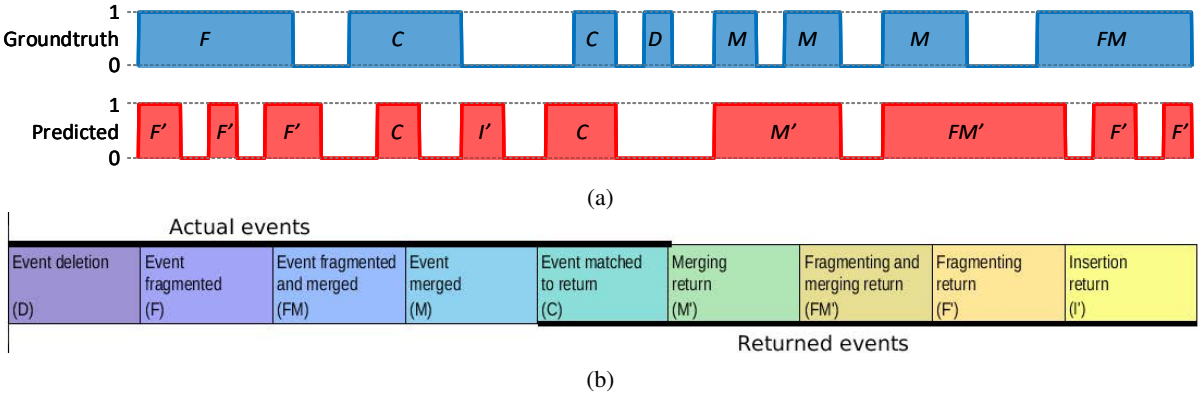


FIGURE 4: (a) Event-based annotation of the ground-truth and predicted label streams. (b) Event Analysis Diagram (EAD). (Adapted from [45])

respect to one another, as insertions (I), deletions (D), fragmentations (F), merges (M), fragmented merges (FM), or correct matches (C), as illustrated by the schematic in Figure 4(a). These event-based measures can be visually summarised in an Event Analysis Diagram (EAD) (Figure 4(b)).

VI. EXPERIMENTAL RESULTS AND DISCUSSION

A. QUALITATIVE EXAMPLE

Audio and video data complement each other. There are times during focsued interaction when visual cues are missing (e.g. when walking side-by-side) or audio cues are missing (e.g. when pausing for thought). Fusion of visual and audio cues facilitates detection of such interaction.

Figure 5(a) shows an example tracker score sequence. Representative frames from that video are shown and labelled (i) - (x). Correct face tracks occur between (iii) to (v) and from (ix) to (x). The true face tracks have greater duration than the false ones, as tends to be the case more generally. At (v), KLT loses track as the person moves out of view but shortly afterwards a new track is generated once the person moves back into view. Likewise, at (x) the track is lost due to full face occlusion but is recovered once the face is detected again after a short while.

Figure 5 (b) shows estimated VAD scores. Voice activity from nearby people is picked up by the detector at (i). At (ii), a focused interaction begins and voice activity is detected but no face is detected until (iii) due to motion blur and distance of the participant from the camera. Another focused interaction begins at (viii); although there is no face present in the field-of-view of the camera at this point, voice activity is detected. Note that even when the tracker is lost at (v) and (x), voice activity is still detected. Due to environmental noise (recording on a windy day), voice activity is falsely detected between (vi) and (viii) albeit with relatively low scores. Automatic doors opening and closing and environmental noise influenced the VAD scores before (i).

B. SINGLE-FRAME CLASS-CONDITIONAL DENSITIES

If the temporal window is set to a single frame ($M = 1$) and the feature set is restricted to VAD and track scores, the resulting two-dimensional feature space can be easily visualised. Figure 6 shows plots of class-conditional densities in this case, estimated from the entire Focused Interaction dataset. The density for the positive class is bimodal, reflecting the natural ebb and flow of conversations occuring during focused interactions with pauses and turn-taking between the subject and conversational partners more distant from the sensor. The negative class density shows a spread of track and VAD scores with a clear peak at low VAD and track score.

C. STANDARD EVALUATION

We report results using several different combinations of feature sets and classifiers. The term *TV-SVM* denotes the use of an SVM classifier with tracker (T) and VAD (V) scores, for example, whereas *FHTV-LSTM* denotes an RNN classifier with the complete feature set. F and H denote face detection and face height scores, respectively (Sec. IV-A). We refer to the method described in Sec. IV-E as either Vid_MIT (visual feature-based), Aud_MIT (audio feature-based) or Vid_Aud_MIT (audio-visual feature-based).

We select the window size (M) through experimentation by varying it and observing SVM performance. The tracker score exhibits a low frequency behaviour (Figure 5(a)) and doesn't require integration of many frames for reliable prediction. On the other hand, the VAD score exhibits large fluctuations with relatively high frequencies during conversation (Figure 5(b)) and needs the integration of adjacent frames for reliable prediction. We found that the performance of TV-SVM remains stable when selecting window sizes between 25 to 50 frames. Therefore, in subsequent experiments we fixed M to 50. Note that this window duration is sufficient to span many of the gaps in voice activity that occur and which result in short intervals of low VAD score.

Table 3 and Figure 7 report framewise measures. From Figure 7, we observed that the performances of visual-only methods were comparatively poor with AUCs of 0.80 and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2850284, IEEE Access

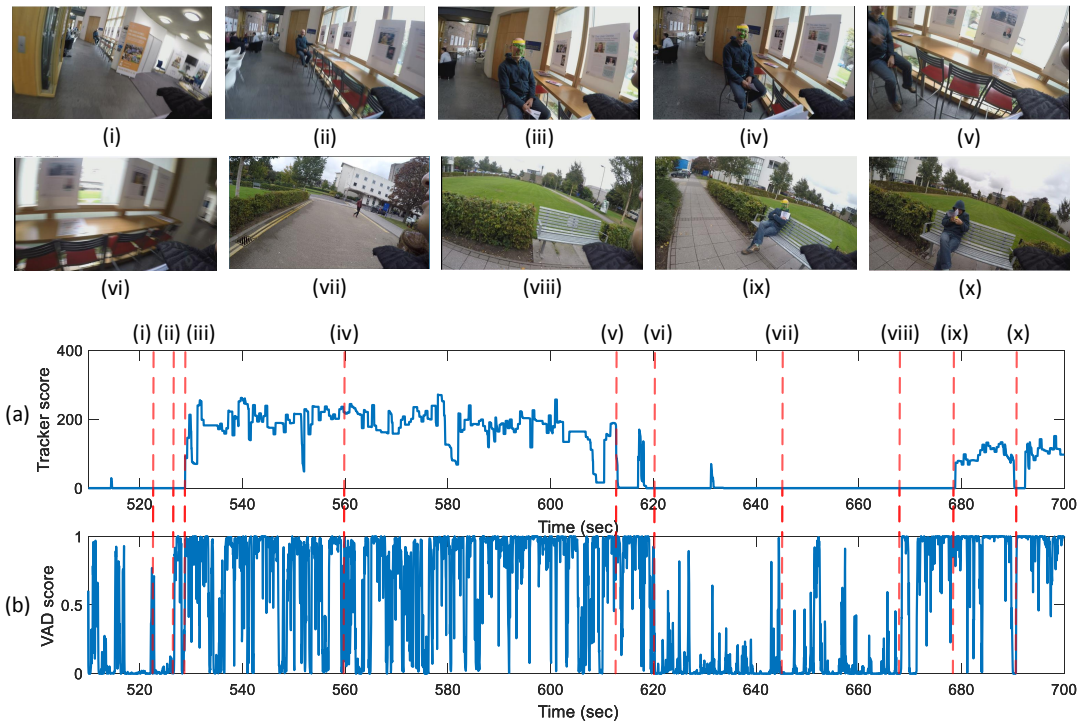
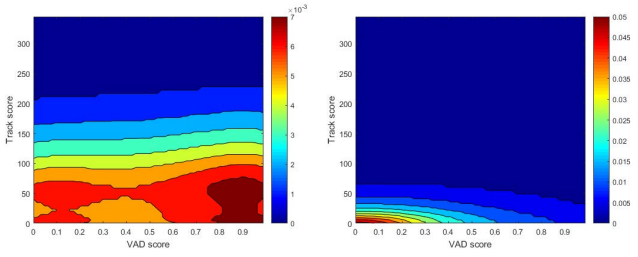


FIGURE 5: (a) Tracker scores and (b) VAD scores from an example sequence. Video frames at times indicated with vertical dashed lines are shown in (i)-(x). A focused interaction starts at (iii) and ends at (vi). Another focused interaction starts at (viii).



(a) (b)

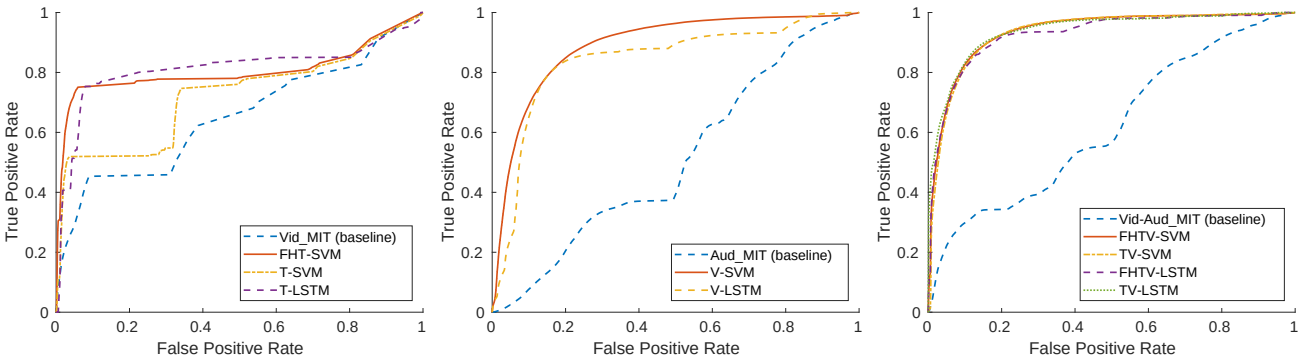
FIGURE 6: Class-conditional densities estimated from tracker and VAD scores at every frame. (a) Density plot for the positive class (focused interaction). (b) Density plot for the negative class (no focused interaction). (Best viewed in colour)

0.72 for FHT-SVM and T-SVM, respectively. The baseline Vid_MIT AUC (0.65) was the lowest among the different video-based feature sets and T-LSTM AUC (0.81) was the highest. Vid_MIT performance was low as it only considers spatial visual features which are not always present due to missed face detection. The use of VAD alone (V-SVM) performed better than the visual-only feature set giving an AUC of 0.89 suggesting voice activity provides a strong cue for the presence or absence of a focused interaction. The baseline Aud_MIT, on the other hand, poorly performed with an AUC of 0.51 due to the use of only basic statistical features compared to VAD that incorporated spectro-temporal behaviour of the human voice. The features in Aud_MIT give the representation of sound in the audio stream and are not capable of differentiating between human voice and environmental noise (sounds from the background).

The best focused interaction detection results were achieved using audio-video fusion: FHTV-SVM and TV-SVM had AUCs of 0.94 and 0.93, respectively and FHTV-LSTM and TV-LSTM had AUCs of 0.93 and 0.94. This suggests that audio-visual fusion is beneficial for focused interaction detection. The baseline Vid_Aud_MIT AUC (0.62) was the lowest because of the poor performance of the audio and visual features used in the method. The F1-scores reinforce this with the best F1 of 0.94 obtained by TV-SVM, followed by FHTV-SVM, FHTV-LSTM, TV-LSTM and V-SVM with F1-scores of 0.93, 0.93, 0.93 and 0.91, respectively. Audio and visual features such as VAD and face track scores complement each other and together provide an effective feature set; addition of face detection score and face height did not help. The presence of a reliable face track and voice activity together provide strong evidence for a focused interaction. The lowest EER for a visual-only feature set was 0.20 for T-LSTM, for audio-only it was 0.18 for both V-SVM and V-LSTM, and for audio-visual it was 0.13 for both TV-SVM and TV-LSTM. This shows that EER with the TV feature set is lower by 5% and 7% than with V and T alone.

We compared results obtained using the SVM and LSTM methods (as detailed in Sec. IV-D). The LSTM method did not yield any significant improvement over the SVM.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2850284, IEEE Access



(a) (b) (c)

FIGURE 7: ROC curves for (a) visual, (b) audio, and (c) audio-visual features.

TABLE 3: Framewise classification measures pooled over validation folds. Key: $\mathcal{P}$ - precision; $\mathcal{R}$ - recall; $\mathcal{F}$ - F1-score; $\mathcal{O}$ - fall-out; AUC - area under curve; EER - equal error rate.

Feature set type	Method	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}$	$\mathcal{O}$	AUC	EER
Visual only	Vid_MIT (baseline)	0.7753	**0.9119**	0.8380	0.8775	0.6508	0.3832
	FHT-SVM	0.7937	0.8257	0.8094	0.7120	0.8038	0.2276
	T-SVM	0.7781	0.8483	0.8117	0.8022	0.7215	0.3265
	T-LSTM	**0.8679**	0.8260	**0.8465**	**0.4140**	**0.8124**	**0.2042**
Audio only	Aud_MIT (baseline)	0.7685	**1.0000**	0.8691	1.0000	0.5109	0.5211
	V-SVM	0.9101	0.9139	**0.9120**	0.2996	**0.8899**	**0.1768**
	V-LSTM	**0.9320**	0.8391	0.8831	**0.2016**	0.8425	0.1784
Visual + Audio	Vid-Aud_MIT (baseline)	0.7925	0.9077	0.8462	0.7888	0.6188	0.4498
	FHTV-SVM	0.9376	0.9275	0.9325	0.2049	0.9358	0.1303
	FHTV-LSTM	0.9412	0.9066	0.9236	0.1864	0.9270	0.1382
	TV-SVM	0.9351	**0.9387**	**0.9353**	0.2155	0.9332	0.1296
	TV-LSTM	**0.9486**	0.9061	0.9269	**0.1618**	**0.9389**	**0.1259**

The AUCs for FHTV-LSTM and TV-LSTM were 0.93 and 0.94, and their F1-scores were 0.92 and 0.93, respectively. Comparing these values and the ROC curves in Figure 7 with those of FHTV-SVM and TV-SVM, we can conclude that the results from these two classifiers are similar. Considering the successful TV feature set, we observe by visualising the predicted stream against ground-truth stream that the errors in SVM-based methods mainly occur due to the fluctuating response of the predicted stream at focused interaction labels. This is because of the varying response of the features at focused interaction (e.g. varying point trackers due to movements during conversation, highs and lows of voice). The LSTM-based method, on the other hand, overcame such fluctuating responses but resulted in delayed detections. The extended measures (discussed below) further helped in analysing these errors.

As supplementary material, we provide example videos visualising the predicted stream for TV-SVM. These videos are short clips extracted from a 52 mins long video stream and highlight the challenges associated with online focused interaction detection in continuous egocentric videos.

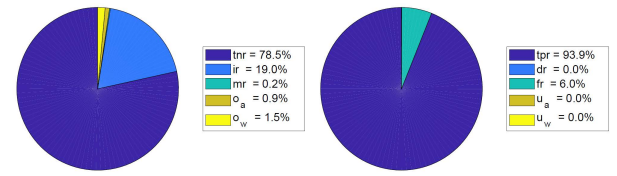
D. EXTENDED EVALUATION

To get further insight into the nature of the temporal segmentations produced by the best performing methods, TV-SVM and TV-LSTM, we report extended framewise and event-based measures with and without temporal filtering (Sec. IV-D3).

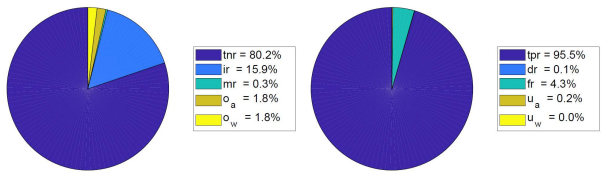
Figure 8 reports results for the unfiltered and filtered TV-SVM. After filtering, the insertion (ir) and fragmentation (fr) errors were reduced by 3% and 1.7% (see Figure 8(a) and (b)); TNR and TPR were also improved to 80.2% and 95.5%, respectively. From the event analysis diagram (Figure 8(c)), it can be observed that only 21 events were correctly predicted out of 64 actual events in unfiltered TV-SVM. A great number of returned events are insertion (1419) and fragmentation (1071). This is because SVM does not consider temporal information beyond the temporal window of $M = 50$ frames. These returned events were generally of short duration. As a result, filtering TV-SVM reduced the insertion returns to 62 and fragmentation returns to 105 (Figure 8(d)). Correct event count also improved to 42 events.

In the case of TV-LSTM, ir and fr were reduced by 2.6% and 1.4% after filtering (see Figure 9(a) and (b)); TPR and TNR were improved to 91.9% and 85%, respectively. From the event analysis diagram (Figure 9(c)), we observe that fragmentation (999) and insertion (310) returns were high but not as high as the TV-SVM (unfiltered). Filtering reduced these errors to 160 and 45, and improved the correct event count to 35 (Figure 9(d)). The filtered fragmentation events

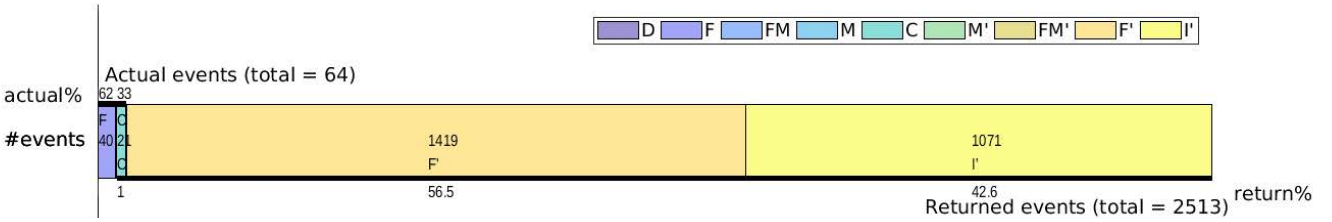
This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2018.2850284, IEEE Access



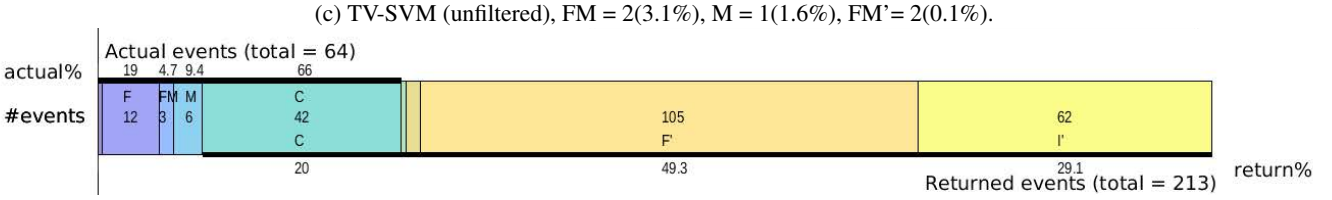
(a) TV-SVM (unfiltered)



(b) TV-SVM (filtered)

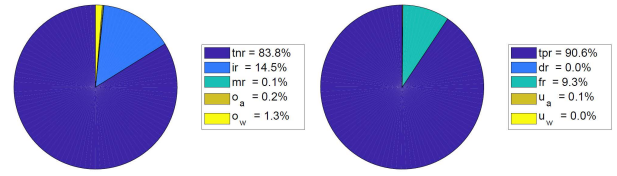


(c) TV-SVM (unfiltered), FM = 2(3.1%), M = 1(1.6%), FM'= 2(0.1%).

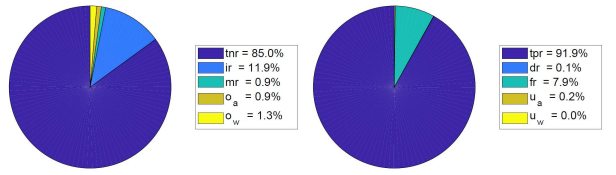


(d) TV-SVM (filtered), D = 1(1.6%), M' = 1(0.5%), FM' = 3(1.4%).

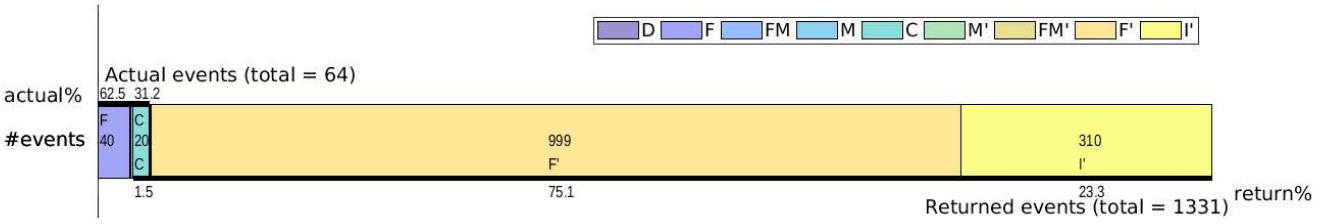
FIGURE 8: Extended framewise evaluation for (a) TV-SVM (unfiltered) and (b) TV-SVM (filtered). Event-based evaluation for (c) TV-SVM (unfiltered) and (d) TV-SVM (filtered). Filtering helps in reducing the fragmentation and insertion errors. Best viewed in colour.



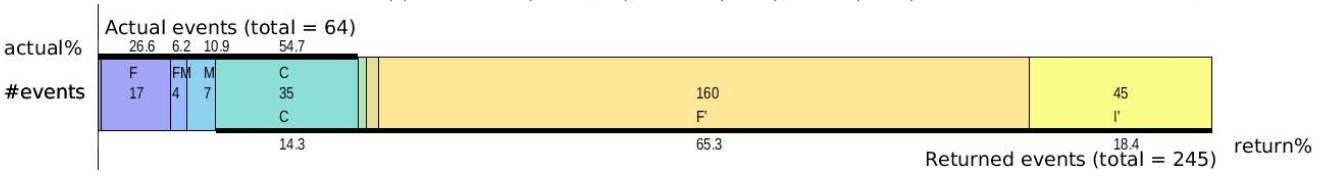
(a) TV-LSTM (unfiltered)



(b) TV-LSTM (filtered)



(c) TV-LSTM (unfiltered), FM = 1(1.6%), M = 3(4.7%).



(d) TV-LSTM (filtered), D = 1(1.6%), M' = 0.8(2.0%), FM' = 3(1.2%).

FIGURE 9: Extended framewise evaluation for (a) TV-LSTM (unfiltered) and (b) TV-LSTM (filtered). Event-based evaluation for (c) TV-LSTM (unfiltered) and (d) TV-LSTM (filtered). Filtering helps in reducing the fragmentation and insertion errors. Best viewed in colour.

for TV-LSTM are higher than TV-SVM (filtered). This was because some fragmentation events in the case of TV-LSTM were of long duration; filtering was unable to correct such errors.

Overall, results from both TV-SVM and TV-LSTM remained comparable even after filtering. This extended evaluation protocol allowed better understanding and visualisation of the classification errors in sequential data. The analysis

highlighted additional challenges of this dataset that include scenarios with background voices (from passing individuals), relatively long pauses during a conversation especially when the conversational partner is not facing the camera wearer, and varied lighting conditions.

VII. CONCLUSION

We have contributed the Focused Interaction dataset capturing everyday interactions from an egocentric perspective in varying locations and environmental conditions. One camera wearer performed all the recordings in this dataset. We presented and evaluated methods for the online detection of focused interaction. In contrast to methods for detection of social interaction that classify video clips we perform continuous segmentation. We processed both audio and visual data streams to obtain audio-visual feature sets. Temporal segmentation of focused interactions was achieved via classification using either SVMs or LSTM recurrent neural networks. Evaluation using various feature sets was performed in terms of framewise and event-based measures and comparison was made with a baseline method. It was shown that integrating audio cues with visual cues improved the performance of focused interaction detection over the use of audio or video alone. The SVM-based method gave more missed-detection intervals of shorter duration compared to the LSTM-based method which gave longer false-detection intervals. The proposed method, the new dataset, and the evaluation protocol provide a benchmark for future research on focused interaction detection.

In the future, larger data sets with multiple subjects could be captured to extend the scope of this work and allow testing of generalisation across subjects. We plan to extend this work to identify conversational partners and scene-specific information during interactions in order to enhance assistive technologies for non-speaking people. Other application domains include behaviour understanding in care settings and evidence management for law enforcement.

ACKNOWLEGEMENTS

The authors are grateful to Annalu Waller, the ACE-LP project team and members of the CVIP group (University of Dundee) for useful discussions and assistance with dataset collection.

REFERENCES

[1] E. Goffman, Encounters: two studies in the sociology of interaction, ser. The advanced studies in sociology series. Bobbs-Merrill, 1961.

[2] R. J. Rummel, Understanding Conflict and War: Vol. 2: The Conflict Helix: Chap 9: Social Behavior and Interaction. Beverly Hills: Sage, 1976.

[3] S. Coradeschi, A. Cesta, G. Cortellessa, L. Coraci, J. GonzÃąlez-JimÃl'nez, L. Karlsson, F. Furfari, A. Loutfi, A. Orlandini, and F. Palumbo, "Giraffplus: Combining social interaction and long term monitoring for promoting independent living," in 6th International Conference on Human System Interaction (HSI). IEEE, 2013, pp. 578–585.

[4] T. Choudhury and A. Pentland, "Sensing and modeling human networks using the sociometer," in 7th IEEE International Symposium on Wearable Computers. IEEE, 2003, pp. 216–222.

[5] S. J. Boutell, C. G. Paulson-Ellis, H. D. S. Martin, A. H. Chisholm, R. A. Iddon, and R. McBride, "Image recording apparatus with slidable concealing cover," Jul. 15 2014, US Patent 8,780,205.

[6] A. Braga, J. R. Coldren Jr, W. Sousa, D. Rodriguez, and O. Alper, "The benefits of body-worn cameras: new findings from a randomized controlled trial at the Las Vegas Metropolitan Police," Washington DC: National Institute of Justice, 2017.

[7] J. Maskaly, C. Donner, W. G. Jennings, B. Ariel, and A. Sutherland, "The effects of body-worn cameras (BWCs) on police and citizen outcomes: A state-of-the-art review," Policing: An International Journal of Police Strategies & Management, vol. 40, no. 4, pp. 672–688, 2017.

[8] S. Bano, J. Zhang, and S. J. McKenna, "Finding time together: Detection and classification of focused interaction in egocentric video," in IEEE International Conference on Computer Vision Workshops (ICCVW). IEEE, Oct 2017, pp. 2322–2330.

[9] D. S. Hayden, "Wearable-assisted social interaction as assistive technology for the blind," Master's thesis, Massachusetts Institute of Technology, 2014.

[10] J. Ramirez, J. M. Górriz, and J. C. Segura, Voice activity detection. fundamentals and speech recognition system robustness. INTECH Open Access Publisher NewYork, 2007.

[11] S. Hizlisoy and Z. Tufekci, "Noise robust speech recognition using parallel model compensation and voice activity detection methods," in 5th International Conference on Electronic Devices, Systems and Applications (ICEDSA). IEEE, 2016, pp. 1–4.

[12] S. Graf, T. Herbig, M. Buck, and G. Schmidt, "Features for voice activity detection: a comparative analysis," EURASIP Journal on Advances in Signal Processing, vol. 91, 2015.

[13] M. van Segbroeck, A. Tsiartas, and S. Narayanan, "A robust frontend for VAD: exploiting contextual, discriminative and spectral cues of human voice," in Interspeech, 2013, pp. 704–708.

[14] M.-W. Mak and H.-B. Yu, "A study of voice activity detection techniques for NIST speaker recognition evaluations," Computer Speech & Language, vol. 28, no. 1, pp. 295–313, 2014.

[15] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009, pp. 935–942.

[16] 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 British Machine Vision Conference, vol. 2, 2011, pp. 1–12.

[17] X. Alameda-Pineda, Y. Yan, E. Ricci, O. Lanz, and N. Sebe, "Analyzing free-standing conversational groups: A multimodal approach," in 23rd ACM International Conference on Multimedia. ACM, 2015, pp. 5–14.

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

[19] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012, pp. 1346–1353.

[20] D. Damen, T. Leelasawassuk, and W. Mayol-Cuevas, "You-do, I-learn: Egocentric unsupervised discovery of objects and their modes of interaction towards video-based guidance," Computer Vision and Image Understanding, vol. 149, pp. 98–112, 2016.

[21] M. Ryoo and L. Matthies, "First-person activity recognition: Feature, temporal structure, and prediction," International Journal of Computer Vision, vol. 119, no. 3, pp. 307–328, 2016.

[22] Y. Poleg, C. Arora, and S. Peleg, "Temporal segmentation of egocentric videos," in IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2014, pp. 2537–2544.

[23] M. Dimiccoli, M. Bolaños, E. Talavera, M. Aghaei, S. G. Nikolov, and P. Radeva, "SR-clustering: Semantic regularized clustering for egocentric photo streams segmentation," Computer Vision and Image Understanding, vol. 155, pp. 55–69, 2017.

[24] A. Fathi, J. K. Hodgins, and J. M. Rehg, "Social interactions: A first-person perspective," in IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012, pp. 1226–1233.

[25] S. Alletto, G. Serra, S. Calderara, and R. Cucchiara, "Understanding social relationships in egocentric vision," Pattern Recognition, vol. 48, no. 12, pp. 4082–4096, 2015.

[26] M. Aghaei, M. Dimiccoli, and P. Radeva, "With whom do I interact? Detecting social interactions in egocentric photo-streams," in IEEE International Conference on Pattern Recognition (ICPR). IEEE, 2016, pp. 2959 –2964.

[27] Y. Yan, E. Ricci, G. Liu, and N. Sebe, "Egocentric daily activity recognition via multitask clustering," IEEE Transactions on Image Processing, vol. 24, no. 10, pp. 2984–2995, 2015.

[28] X. Alameda-Pineda, V. Khalidov, R. Horaud, and F. Forbes, "Finding audio-visual events in informal social gatherings," in 13th International Conference on Multimodal Interfaces. ACM, 2011, pp. 247–254.
[29] I. D. Gebru, X. Alameda-Pineda, R. Horaud, and F. Forbes, "Audio-visual speaker localization via weighted clustering," in IEEE International Workshop on Machine Learning for Signal Processing. IEEE, 2014, pp. 1–6.
[30] D. Chen, J. Yang, R. Malkin, and H. D. Wactlar, "Detecting social interactions of the elderly in a nursing home environment," ACM Transactions on Multimedia Computing, Communications, and Applications, vol. 3, no. 1, 2007.
[31] K. Mahkonen, J.-K. Kämäräinen, and T. Virtanen, "Lifelog scene change detection using cascades of audio and video detectors," in Asian Conference on Computer Vision. Springer, 2014, pp. 434–444.
[32] H. Hung and B. Kröse, "Detecting F-formations as dominant sets," in 13th International Conference on Multimodal Interfaces. ACM, 2011, pp. 231–238.
[33] M. Aghaei, M. Dimiccoli, C. Canton-Ferrer, and P. Radeva, "Social style characterization from egocentric photo-streams," in IEEE International Conference on Computer Vision Workshops (ICCVW). IEEE, Oct 2017.
[34] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in 18th ACM International Conference on Multimodal Interaction. ACM, 2016, pp. 279–283.
[35] K. G. M. Chathuramali and R. Rodrigo, "Faster human activity recognition with SVM," in International Conference on Advances in ICT for Emerging Regions, Dec 2012, pp. 197–203.
[36] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine," in International Workshop on Ambient Assisted Living. Springer, 2012, pp. 216–223.
[37] N. Y. Hammerla, S. Halloran, and T. Ploetz, "Deep, convolutional, and recurrent models for human activity recognition using wearables," in International Joint Conference on Artificial Intelligence, 2016, pp. 1533–1540.
[38] G. Abebe and A. Cavallaro, "A long short-term memory convolutional neural network for first-person vision activity recognition," in IEEE International Conference on Computer Vision Workshops (ICCVW), Oct 2017, pp. 1339–1346.
[39] S. Yao, S. Hu, Y. Zhao, A. Zhang, and T. Abdelzaher, "Deepsense: A unified deep learning framework for time-series mobile sensing data processing," in 26th International Conference on World Wide Web, 2017, pp. 351–360.
[40] Y.-C. Su and K. Grauman, "Detecting engagement in egocentric video," in European Conference on Computer Vision, 2016, pp. 454 – 471.
[41] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1. IEEE, 2005, pp. 886–893.
[42] B. Benfold and I. Reid, "Stable multi-target tracking in real-time surveillance video," in IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2011, pp. 3457–3464.
[43] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2012, pp. 2879–2886.
[44] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," ACM SigKDD Explorations Newsletter, vol. 12, no. 2, pp. 74–82, 2011.
[45] J. A. Ward, P. Lukowicz, and H. W. Gellersen, "Performance metrics for activity recognition," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 2, no. 1, 2011.

SOPHIA BANO was a Postdoctoral Research Assistant at University of Dundee (UK) during the research described here. Since June 2018 she is a Research Associate at University College London (UK). Previously, She received the Bachelor in Mechatronics Engineering and Master in Electrical Engineering degrees from National University of Sciences and Technology (NUST), Pakistan, in 2005 and 2007, respectively. Following this, in 2011, she received distinction in Erasmus Mundus MSc in Computer Vision and Robotics (VIBOT) programme run jointly by Heriot-Watt University (UK), University of Girona (Spain) and University of Burgundy (France). This was followed by a joint PhD in Interactive and Cognitive Environments (Electronics Engineering) in 2015, under the Erasmus Mundus Fellowship, from Queen Mary University of London (UK) and Technical University of Catalonia (Spain). Her research interests include computer vision, machine learning and signal processing for multimodal time-series data analysis.

TAMAS SUVEGES is currently a PhD student in Computing, School of Science and Engineering, University of Dundee, UK. His research interests include deep learning for recognition and retrieval.

JIANGUO ZHANG is currently a Reader in Computing, School of Science and Engineering, University of Dundee, UK. He received a PhD in National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, 2002. His research interests include object recognition, image processing, medical image analysis, machine learning and computer vision for assistive technology. He is a senior member of the IEEE.

STEPHEN J. MCKENNA is Chair of Computer Vision and Head of Computing Research at the School of Science and Engineering, University of Dundee. He received a B.Sc. degree in Computer Science from the University of Edinburgh (1990) and Ph.D. in Medical Image Analysis from the University of Dundee (1994). His research interests include biomedical image analysis, computer vision and machine learning.