Gestures In-The-Wild

Detecting Conversational Hand Gestures in Crowded Scenes Using a Multimodal Fusion of Bags of Video Trajectories and Body Worn Acceleration

Cabrera Quiros, Laura; Tax, David M.J.; Hung, Hayley

DOI
10.1109/TMM.2019.2922122

Publication date
2020

Document Version
Final published version

Published in
IEEE Transactions on Multimedia

Gestures In-The-Wild: Detecting Conversational Hand Gestures in Crowded Scenes Using a Multimodal Fusion of Bags of Video Trajectories and Body Worn Acceleration

Laura Cabrera-Quiros, David M. J. Tax, *Member, IEEE*, and Hayley Hung, *Member, IEEE*

***Abstract*—This paper addresses the detection of hand gestures during free-standing conversations in crowded mingle scenarios. Unlike the scenarios of the previous works in gesture detection and recognition, crowded mingle scenes have additional challenges such as cross-contamination between subjects, strong occlusions, and nonstationary backgrounds. This makes them more complex to analyze using computer vision techniques alone. We propose a multimodal approach using video and wearable acceleration data recorded via smart badges hung around the neck. In the video modality, we propose to treat noisy dense trajectories as bags-of-trajectories. For a given bag, we can have good trajectories corresponding to the subject, and bad trajectories due for instance to cross-contamination. However, we hypothesize that for a given class, it should be possible to learn trajectories that are discriminative while ignoring noisy trajectories. We do this by exploiting multiple instance learning via embedded instance selection as our multiple instance learning approach. This technique also allows us to identify which instances contribute more to the classification. By fusing the decisions of the classifiers from the video and wearable acceleration modalities, we show improvements over the unimodal approaches with an AUC of 0.69. We also present a static analysis and a dynamic analysis to assess the impact of noisy data on the fused detection results, showing that the moments of high occlusion in the video are compensated by the information from the wearables. Finally, we applied our method to detect speaking status, leveraging the close relationship found in the literature between hand gestures and speech.**

***Index Terms*—Hand gestures, crowded mingles, dense trajectories, multiple instance learning, MILES, wearable acceleration.**

Manuscript received August 20, 2018; revised March 19, 2019 and May 1, 2019; accepted May 2, 2019. Date of publication June 11, 2019; date of current version December 31, 2019. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jingdong Wang. *(Corresponding author: Laura Cabrera-Quiros.)*

L. Cabrera-Quiros is with the Department of Intelligent Systems, Delft University of Technology, Delft, The Netherlands, and also with the Escuela de Ingeniería Electrónica, Instituto Tecnológico de Costa Rica, Costa Rica (e-mail: l.c.cabreraquiros@tudelft.nl).

D. M. J. Tax and H. Hung are with the Department of Intelligent Systems, Delft University of Technology, Delft, The Netherlands (e-mail: d.m.j.tax@tudelft.nl; h.hung@tudelft.nl).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2019.2922122

I. Introduction

HAND gestures constitute one of the key elements of face to face interactions. As described by Adam Kendon [1](1): *"Willingly or not, humans (...) [communicate] their intentions, interests, feelings and ideas by means of visible bodily action."* During a conversation there is a high probability of observing conversational hand gestures, and their analysis can provide further insights about the interaction itself [2], [3].

However, current efforts on the automatic detection and recognition of gestures does not focus on the gestures that we all perform on a daily basis, which are mostly inherently conversational as described by [1] and [2]. Instead, most works in the Multimedia, Computer Vision and Human Computer Interaction (HCI) communities are focused on scenarios where the person is only performing symbolic gestures[1] that are clearly visible. For example, over the last five years the gesture recognition Chalearn challenge [4], [5] has provided over 40000 videos of one person at a time performing sign language gestures in front of a Kinect (see Figure 1(a)), either trimmed or with consecutive gestures. Several works have used these datasets to address the problem of gesture recognition under these conditions [6], [7].

Unfortunately, this does not reflect the majority of real life situations where gestures are used. These works, although interesting for certain applications (e.g. HCI), only address a subset of the wide variety of gestures a human can perform, and this subset has a consistent and discriminative pattern [3] (e.g. 'hello' in sign language is always the same). Also, they present rather stationary backgrounds with a single subject, without any cross-contamination between subjects.

In contrast, datasets for social interaction analysis *in-the-wild* should maintain ecological validity. Hence, these types of scenarios have a crowded nature as the people come together to form conversational groups. To better capture these events, a top view is preferred (see Figure 1(c)). Side or elevated view tend to have higher amounts of occlusions, particularly for those people away from the camera (see Figure 1(b)).

The scenarios studied in this paper are *crowded mingle events*, where people are inherently encouraged to interact in a real setting (e.g. parties). Thus, they provide a perfect example of the use

[1]Symbolic gestures are those with a specific meaning (e.g. thumbs up).

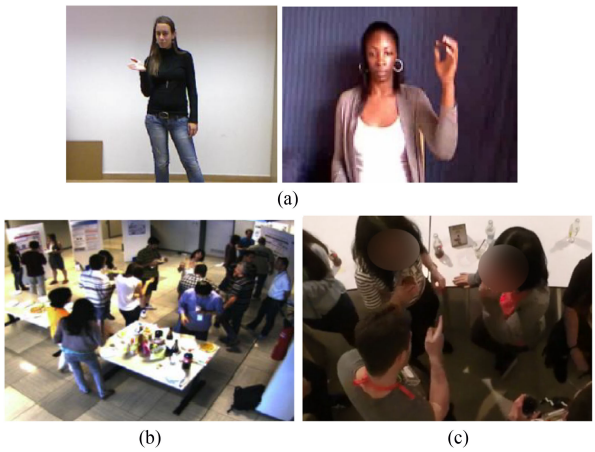


Fig. 1. Different scenarios for hand gesture detection. (a) Symbolic gesture [4], [5], Conversational gestures: (b) Salsa Dataset [8] and (c) Our scenario.

of hand gestures during real conversations within a social context in an *in-the-wild* scenario. Nonetheless, we hypothesize that our method can be applied in other cases with related contexts.

From the visual modality perspective, mingle scenarios have four main differences when compared to the symbolic gesture scenarios: 1) cross-contamination between subjects, where using a bounding box as it is done by most works in object/person detection could include two or more people in one person's box; 2) strong occlusions, making it impossible to see the subjects in some cases; 3) strong changes in appearance for the same subject; and 4) non-stationary backgrounds, which are affected by the position of the subjects, lighting conditions or shadows. A visual depiction of these challenges is shown in Figure 2. Thus, the detection of gestures in such scenarios must deal with the presence of noisy data.

Fortunately, these challenges can be addressed using multiple modalities, as shown by previous work [9], [10]. Moreover, works on free-standing conversational groups have shown that wearable sensing alternatives can provide additional information when learning with video [11]. Thus, each modality can provide different information to understand the event, relying on one modality when the other is missing or leveraging their complementarity.

In this paper we detect conversational hand gestures in crowded and strongly occluded scenarios using dense trajectories from video and wearable acceleration.

Our main contributions are: (i) unlike previous works, we addressed the detection of conversational hand gesture *in-the-wild*, using a dataset collected during a real mingle event with strong interpersonal occlusions, (ii) we propose to use a multiple instance learning approach (MILES [12]) representing gestures as bags of trajectories to overcome subject cross-contamination and to become robust against noisy backgrounds in video, (iii) we leverage the MILES instance classification capability to analyze which dense trajectories (in time and space) are more representative for a gesture in video, (iv) we combine video and wearable acceleration in a decision-level manner leveraging the complementarity between modalities, particularly for cases where occlusions in video are too strong to have a clear view of the person performing the gesture, showing improvements over unimodal approaches; (v) we analyze the impact of noisy data (e.g. strong occlusions) of the participants on the overall performance, both static and dynamically in time; and finally, (vi) we use our method to detect speaking status, leveraging the relationship between people's gestures and speech [1], [2].

To the best of our knowledge we are the first to address the problem of gesture detection in crowded scenes, emphasizing the importance of the social context of the gestures and its impact in the challenges for the detection of gestures.

The rest of the paper is divided as follows. Section II presents related efforts about gesture detection in-the-wild. The dataset used is described in Section III. Section IV gives a detailed description of our approach, the feature extraction for video and the wearable sensors, and the process of decision fusion of these 2 modalities. Our experiments are presented in Section V, and its discussion in Section VI. Finally, we conclude in Section VII.

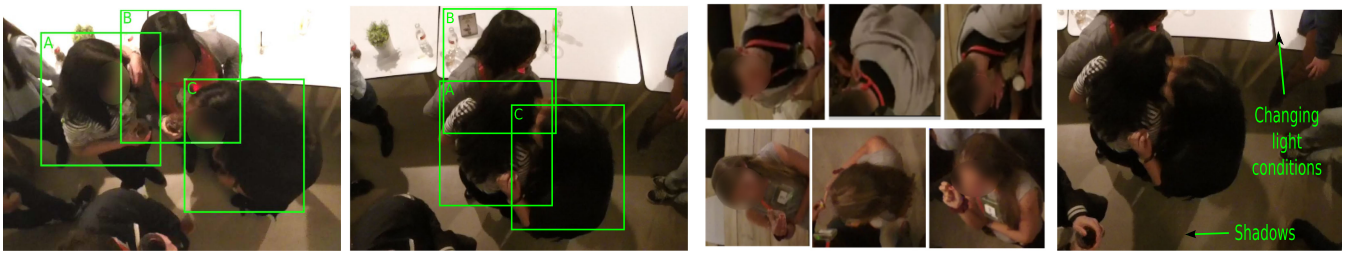
II. Related Work

Most work on the detection and recognition of gestures focuses on cases where there is a clear view of the person performing a symbolic gesture, generally from the front. For these works the process is quite similar: 1) pose estimation or use of its skeleton if available, and 2) gesture detection.

For instance, one of the datasets for gesture recognition for the Chalearn 2016 [5] was the LAP Continous Gesture Dataset (ConGD). This dataset consists of over 45000 RGB-D gestures within over 22000 RGB-D videos. Each video may represent one or more gestures, and there are 249 gestures labels performed by 21 different individuals. For detecting the gestures segments within the video, the teams with the best performances use sliding windows of video as input for a convolutional 3D neural network [13], or finding the start and end frames of each gesture using quantity of movement (QOM), by assuming that all gestures starts from a similar and clear pose [14].

Some works also include a more strict hand segmentation step in this pipeline. Chai *et al.* [15] used hand detection and their position for a temporal segmentation, achieving the best performance for the Chalearn gesture detection and recognition of 2016 [5]. Ren *et al.* [16] use hand detection and segmentation for shape representation, in order to recognize hand gestures using a Kinect. Something similar was proposed by Wang *et al.* [17] using superpixels, and Liang *et al.* [18] proposed a parsing scheme for hand representation on 3D, also based on superpixels. Alon *et al.* [19] addressed the problem of spatiotemporal gesture segmentation for the American Sing Language (ASL). To do so, they presented an unified a framework for simultaneously performing spatial segmentation, temporal segmentation, and recognition, which also starts with a hand detection based on motion detectors. Same as the works using the Chalearn dataset, their dataset consist of videos of people gesturing in front of a camera, and each video may represent one or more gestures.

These works, although relevant for its specific goal such as interacting with a computer, do not address the same inherent problem as our work and most do not share the same challenges (eg. strong occlusions due to natural interactions).



(a) (b) (c) (d)

Fig. 2. Challenges while analyzing mingle scenarios with video (see better in color). (a) Interpersonal cross-contamination, (b) strong interpersonal (A to B) and intrapersonal (A to herself) occlusions, (c) strong appearances variations for two subjects, and (d) Non-stationary background.

For a more comprehensive review of the domain, please refer to any of the past Chalearn Challenges about gesture recognition [4], [5] and to [20] for a review on gesture recognition for HCI applications.

In contrast to the works above, here we focus on related works with scenarios similar to ours. Xiong and Quek [21] presented their work on the analysis of gestures during conversations for the analysis of the frequency of gestures. To do so, they applied a windowed Fourier transform and wavelet transform to detect and extract gesticulatory oscillations. Although oriented to conversational gestures, this work is based on a dataset which has a rather clear side view of the speaker and only 30 seconds of video [22].

Similarly, Marco-Ramiro *et al.* [23], [24] addressed the detection of conversational gestures during seated encounters. Their first work focused on the detection of upper body monocular motion including hands using an approximate 3D upper body pose, while the latter used such features to look for adaptors (meaning unintentional gestures, generally performed while fidgety) or beat gestures in the context of an interview. Also, Cerekovic *et al.* [25] detected the rapport between people and virtual agents using as one of their features the hand gesture activity of the people while interacting with the agent.

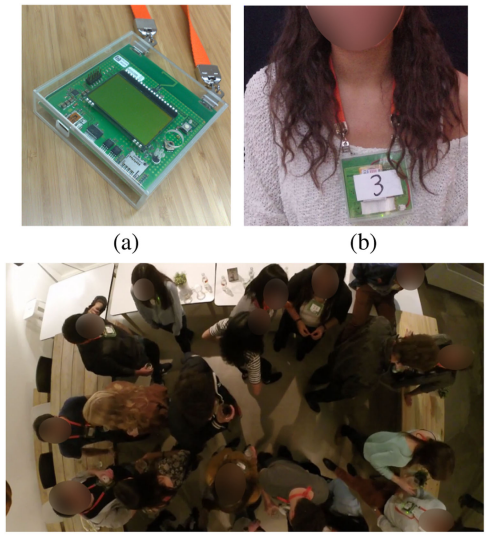
These efforts addressed the wide range of human gestures during conversations. Nonetheless, they used a rather clear front view of the participants while interacting, so they do not present the same additional challenges regarding the visual perspective (e.g. cross-contamination) as our mingle scenario.

The closest to our work, regarding the use of multiple instance learning for detecting gestures, was presented by Ali and Shah [26] and Yi and Lin [27]. Both works showed methods based on multiple instances for general activity recognition. Nevertheless, they did so in the KTH, UCF sports, Youtube and Hollywood action data sets, which do not present crowded scenes and do not include hand conversational gestures as part of their classes.

To the best of our knowledge we are the first to address the problem of gesture detection during crowded scenes.

III. Crowded Mingle Scenario

We use the MatchNMingle dataset [28], a multimodal resource for the analysis of social interactions in the wild.[2] This dataset provides information for up to 70 people while mingling freely for 30 minutes. Only 10 minutes are used in this work.



(a) (b)

(c)

Fig. 3. (a) Wearable device used by participants. (b) Wearing method for device. (c) Our crowded mingle scenario, with multiple conversations, and strong interpersonal cross-contamination and occlusions [28].

During one of three different day events, participants were part of a speed date event, each followed by a mingle session. As we focus on detection of gestures in-the-wild during standing multi-party conversations, we will only use the mingle part of MatchNMingle but we hypothesize that our insights here can be also applied to a seated scenario, as the speed dates.

For the mingle session, the participants were not instructed in any way, and they can move freely through the mingle area or leave it at will (e.g. go to the bathroom). They can also order food or drinks during the entire event. Thus, their gestures are inherent to the social interactions they are having with other participants or with members of the staff (e.g. waiters).

Each participant wore a smart badge hung around the neck (see Figure 3(a-b)) recording triaxial acceleration at 20Hz during the entire event. In addition, video was recorded at 20 FPS from above (see snapshot on Figure 3(c)).

Finally, the dataset also provides the manual annotations of the social actions (e.g. speaking, hand gestures) for all participants and the ground truth for their positions in the image. These

[2]Dataset is openly available under an EULA and can be found in http://matchmakers.ewi.tudelft.nl/matchnmingle/pmwiki/.

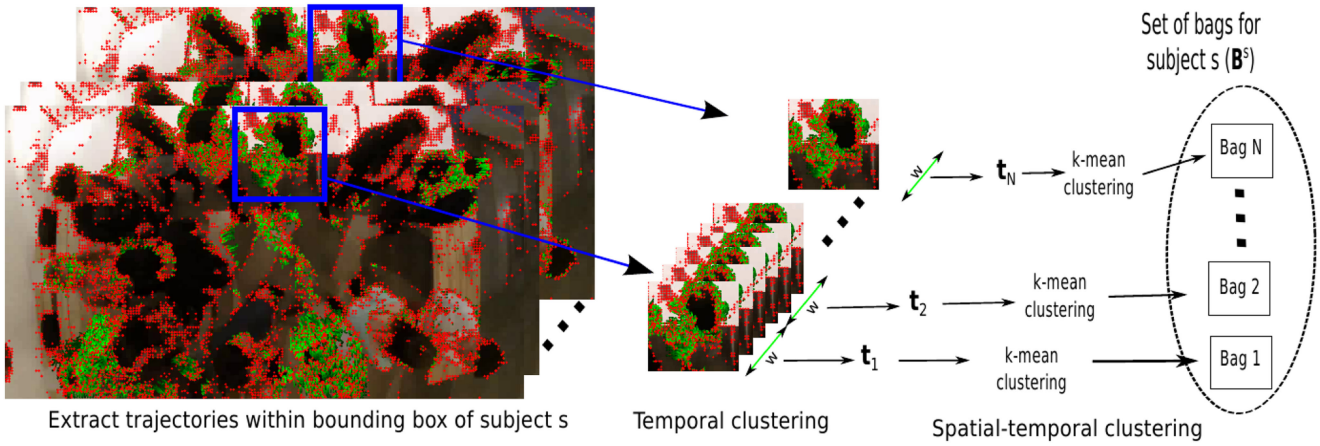


Fig. 4. Process of clustering in space and time to create bags of trajectories.

were obtained from trained annotators, with an inter-annotator agreement coefficient of 0.61 for hand gestures and 0.55 for speaking, using the kappa-Fleiss coefficient.

IV. Proposed Approach

We apply a window-based approach that identifies if a time interval (window) of length w contains a gesture or not. To do so, it uses as inputs 1) RGB video and 2) the triaxial acceleration on the wearable device of each participant.

For the video modality, we explain in Section IV-A the process of extracting and clustering dense trajectories that are subsequently used as bags of instances in a multiple instance learning classification. In Section IV-B we explain the extraction of features and classification for the wearable modality (devices). Finally, in Section IV-C we explain the process of fusing both modalities in the decision level [29], by using the posterior probabilities of the unimodal classifiers (one per modality) as input to a third classifier.

A. Video Classification

Figure 4 summarizes the process of clustering trajectories in space and time to create our *bags of trajectories* in video for each subject. This process consists of the following steps:

1) Extraction of Dense Trajectories: Firstly, we extract trajectories using the method of dense trajectories proposed by Wang *et al.* [30]. These have proven to be an efficient representation for human activity recognition [30]–[32]. The dense trajectories are extracted for the entire frame using a length L of 20 frames.

Using the bounding boxes for each participant on each frame we create a voxel following the participant over time (see bottom of Figure 5). Thus, we reduce the number of trajectories to those around or from each participant by selecting only those inside this voxel. This selection also accounts for trajectories that start outside the voxel but enter it, and those that start within the voxel and drift out.

For bounding box extraction one can use any existing tool for this purpose [33], [34], however we use the ground truth annotations to avoid further contamination. Our method uses

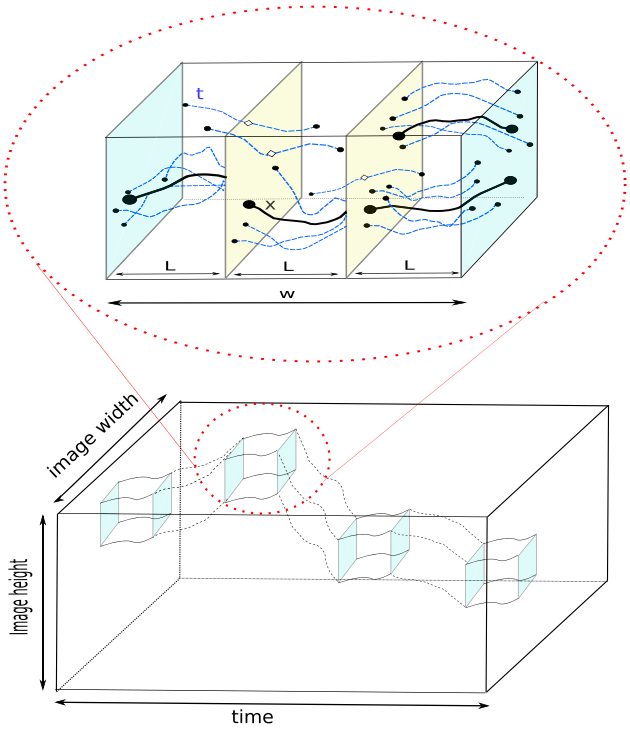


Fig. 5. Clustering of trajectories for a bag.

bags of video features, so it should be robust against small shifts of the bounding boxes. Nonetheless, we leave the analysis of the impact of the errors in the detection and tracking of people on the overall gesture detection for future research, as this lies outside of the scope of this paper.

Notice that, due to the crowdedness of the scene, bounding boxes of different subjects can heavily overlap. A bounding box can therefore contain trajectories of both the subject of interest and of 'background' subjects. Fortunately, our multiple instance learning (MIL) approach will account for this duality.

2) Selection of Trajectories for a Bag: First, define $\mathbf{B}^s$ as the set of positive and negative bags created using the bounding boxes for subject s, where $s = \{1, ..., S\}$ and S is the total number of subjects. A bag from this set is then $\mathbf{B}^s_j$, where

$j = \{1..., N^s\}$, and N^s is the total number of bags possible for subject s.

To create a given window $\mathbf{B}_j^s$ we 1) select the trajectories corresponding to this bag (temporal clustering), and 2) cluster the trajectories within a bag (spatio-temporal clustering). The latter will be explained in the next subsection.

For the temporal clustering, given a sliding window in time and the bounding boxes of subject s, all trajectories that fall in this spatio-temporal box (or voxel) for at least 80% are collected for this bag ($\mathbf{B}_j^s$). This set of trajectories $\mathbf{t}_j^s$ is shown in Figure 5 in blue.

Note that the sliding window does not necessarily have to be the same length L as the trajectories (as presented in Figure 5), or the same shift. Also, the trajectories can start at any point within the sliding window, but will always have a size L. This size is fixed to L to avoid drift, as explained by Wang *et al.* [30]. So, if $w > L$ there will be trajectories in the bag that are only partially within the window.

It is important to emphasize that although the bounding boxes for each subject s were used, **the trajectories inside these boxes do not necessarily belong to subject s**. Instead, they could also represent the background or other subjects. This is the main motivation for using our MIL approach (more in Section IV-A4).

3) Clustering of Trajectories Within a Bag: A final clustering within the bags is important to create less noisy and more representative trajectory prototypes. For practical purposes, it also results in a more efficient memory usage, without losing information. This is a common practice in works using dense trajectories [31], [32], as these descriptors tend to be redundant for similar local-temporal instances.

To cluster the trajectories within a bag, we use k-means clustering. This way, the trajectories for each bag $\mathbf{t}_j^s$ are clustered into the k most representative prototypes for the bag ($\mathbf{x}_{ij}^s$; $i = 1, ..., k$). The trajectories $\mathbf{t}_j^s$ are illustrated in blue in Figure 5. The prototypes $\mathbf{x}_{ij}^s$ are represented in black, and become the instances of our bags ($\mathbf{B}_j^s = \{\mathbf{x}_{ij}^s; i = 1, ..., k\}$).

The label of each bag y_j^s is set using the annotations provided by the dataset, which are made every frame. To select a single value for the bag we use majority voting.

Finally, we create the set of bags $\mathbf{B}^s$ (positive and negative) for subject s by applying the procedure described above for a window, then sliding it and repeating for the entire video, as seen in Figure 5. Thus, each window becomes a bag (see the rightmost part of Figure 4).

4) Multiple Instance Learning: As stated before, due to the crowded nature of our scenes we opt for a multiple instance learning approach using bags of trajectories. Thus, each bag $\mathbf{B}_j^s$ consists of *good* trajectories corresponding to subject s, and *bad or noise* trajectories which could be other subjects or shadows and other background artifacts.

As our Multiple Instance Learning (MIL) approach we use Multiple Instance Learning via Embedded Instance Selection (MILES) [12]. MILES classifies a bag by considering both contributing information (e.g. trajectories of subject s in our case) and opposing information (e.g. trajectories from other subjects or background). It does so by creating a *concept* in an embedded space and comparing all instances to this concept. Instances close to the concept will have a higher contribution (see more on the explanation of Eq. 3).

Thus, unlike other MIL approaches where at least one positive instance in a bag automatically converts it into a positive bag, MILES does not have this restriction. This also allows us to assess the role of individual instances in the classification of a bag (see Section V-B1 for an analysis).

More specifically, MILES maps each bag into a feature space defined by the instances in the training set using bag to instance similarities. The bags are then classify in this space, depending on how close the instances within the bag are to the concept defined by the instances in the training.

Let us define $\mathbf{B} = \{\mathbf{B}^1, \mathbf{B}^2, ..., \mathbf{B}^S\}$, as the set of bags for all participants. $\mathbf{B}_a$ is then a bag of this set $\mathbf{B}$, where $a = \{1..., A\}$ and A is the sum of the total number of bags for all S subjects.

For a given bag $\mathbf{B}_a$ the measure of similarity between this bag and all other instances in the training set[3] (disregarding their bag) is calculated by

$$s(\mathbf{x}^k, \mathbf{B}_a) = \max_b \exp\left(-\frac{||\mathbf{x}_{ab} - \mathbf{x}^k||^2}{\sigma^2}\right) \tag{1}$$

Thus, $s(\mathbf{x}^k, \mathbf{B}_a)$ is the measure of similarity between a concept within the training set (a gesture, in our case) and the bag $\mathbf{B}_a$, which is determined by the closest instance in the bag to the concept.

Using this similarity approach, any bag can be embedded into a similarity space with coordinates $\mathbf{m}(\mathbf{B}_a)$ defined as

$$\mathbf{m}(\mathbf{B}_a) = [s(\mathbf{x}^1, \mathbf{B}_a), s(\mathbf{x}^2, \mathbf{B}_a), ..., s(\mathbf{x}^{n_a}, \mathbf{B}_a)]^T \tag{2}$$

where n_a is the total number of instances in the training set. Applying the mapping in Eq. 2 for a training set of A bags results in the matrix representation of all training bags in the embedded space: $\mathbf{m}(\mathbf{B}) = [\mathbf{m}(\mathbf{B}_1), ..., \mathbf{m}(\mathbf{B}_A)]$.

The creation of this similarity matrix, which is directly dependent of the number of bags (A) and instances in the training set (n_a), can be memory consuming. This is one of the main reasons behind the spatio-temporal clustering within the bags using k-means.

On this representation a (sparse) linear classifier is then trained. The classification of new bags is done by:

$$y = \text{sign}\left(\sum_{k \in I} w_k^* s(\mathbf{x}^k, \mathbf{B}_{new}) + b^*\right) \tag{3}$$

where I is the subset of instances with non-zero weights ($I = \{k : |w_k^*| > 0\}$). Note that instances with contributing information will have positive weights w_k^*, while those with opposing information will have negative weights.

We are also interested in analyzing qualitatively which instances in the bags contributed the most to the MILES classifier. Our intention is to assess whether the instances chosen by the classifier correspond in fact to trajectories of the correct subject, and that trajectories for changes in the background or other subjects are ignored.

[3]The separation of $\mathbf{B}$ into train and test set is addressed in Section V.

For this, we leveraged the instance classification capacity of the MILES algorithm. Thus, for the classification of a given new bag $\mathbf{B}_i$ with instances $\mathbf{x}_{ij}$, $j = 1, ..., n_i$ (where n_i is the total number of instances in the bag), we can define which instances contributed the most for the classification of the bag. To measure this instance level contribution we use the following weight:

$$g(\mathbf{x}_{ij^*}) = \sum_{k \in I_{j^*}} \frac{w_k^* s(\mathbf{x}^k, \mathbf{x}_{ij^*})}{m_k} \tag{4}$$

where $\mathbf{x}^k$ corresponds to the instances in the training set. I_{j^*} corresponds to the subset of all instances in the new bag for which there is a maximum similarity with one of the instances in the training set ($\exp(||\mathbf{x}_{ij} - \mathbf{x}^k||^2/\sigma^2)$) and whose weights are $|w_k^*| > 0$. Finally, m_k is the number of instances in I_{j^*}. Hence, Eq. 4 determines the contribution of $\mathbf{x}_{ij^*}$ on the classification of the bag $\mathbf{B}_i$. For more details about the MILES algorithm, please refer to [12].

B. Wearable Acceleration Classification

Each wearable device (one per subject) recorded the triaxial acceleration at 20 Hz. For each participant, we also calculate the magnitude of the acceleration ($|accel| = \sqrt{x^2 + y^2 + z^2}$); resulting in 4 different time series (x, y, z and $|accel|$) for which we can extract features using a sliding-window approach, similarly to the video. With the triaxial time series we address those movements where the direction is important, whereas with the magnitude we focus on movement in a direction invariant manner.

Then, for each the same sliding windows as defined in the previous section, we extract features that have proven to be efficient to analyze human actions from wearable acceleration [35]. These features are mainly statistical and spectral, where the statistical features focused on mean and variances from each axis and from the magnitude, and spectral using the power spectral density (PSD). All features are concatenated to obtain a 70-dimensional feature vector per window, and then classified using a logistic regressor (see Section V for details).

C. Decision Fusion Classifier

For fusion, we selected a decision-level combining approach [29]. Thus, the approximate posterior probability of the video and wearable classifiers are used as input for a third classifier. We opt for decision fusion instead of early fusion (e.g. concatenate features) as we aim to maintain a constant and fair feature space. This means that, while MILES will map each bag to a embedded space defined by its instances similarities, this can not be applied to the features from the wearable acceleration, thus making an early fusion unfeasible.

Also, as the MILES classifier bases its output on the sum in Eq. 3 instead of a proper probability, we applied Platt scaling [36] to obtain the probability distribution of the classifier from video, as is generally done for similar classifiers such as the SVM.

TABLE I
SUMMARY OF GESTURE DETECTION RESULTS USING UNIMODAL CLASSIFIERS AND THEIR FUSION IN A DECISION-LEVEL. MEAN AUC (± DEVIATION) OF FOLDS

Classifier	Wearable Device	Video (Baseline)	Video (MILES)	Fusion
AUC	0.65 ± 0.08	0.61 ± 0.07	0.67 ± 0.09	$\mathbf{0.69 \pm 0.10}$

V. EXPERIMENTS

We now proceed to evaluate our classifiers, both separately and combined using decision fusion. A summary of the results is presented in Table I. All these values are statistically significant, with $p < 0.01$ when compared to a classifier assigning labels at random. A detailed explanation of each classifier is now presented.

A. Wearable Acceleration Classification

We selected a window size (w) of 60 samples (3 seconds) with no overlap. Empirical tests shown that these values are optimal for our task. For these windows, we extracted for each participant in our dataset (70 in total) the features described in Section IV-B. As a classifier we selected a linear logistic regressor and used a leave-one-subject-out cross-validation strategy.

We obtained a mean AUC of 0.65 ± 0.08 for the 10 minute interval. This result is similar to what has been found in the past for the detection of other social actions using wearable acceleration [37].

B. Video Classification

For each participant we extracted their set of bags of trajectories $\mathbf{B}^s$ following the process described in Section IV-A. Identical to the wearable acceleration, we selected a window size (w) of 60 samples (3 seconds) and no overlap. Hence, for a segment of 10 minutes we obtained a maximum of 200 bags per participant. Some of the participants had less bags, as they left the field of view of the camera for different intervals of time during this interval.

Then, we proceeded to evaluate our MILES approach using leave-one-subject-out. To do so, all bags are extracted for all S participants ($\mathbf{B}^s$, $s = \{1, ..., S\}$). The set of bags for one subject is used for testing while the remaining sets of bags (for $S - 1$ participants) are used for training. This is repeated until all S subjects are used for testing. The mean AUC of the folds (and its deviation) is finally presented.

However, this procedure results in a training set of around 13,500 bags with a total of around 270,000 instances per training fold (exact number depends on the subjects), even when applying the k-means clustering described in Section IV-A. Consequently, the creation of the matrix of distance $\mathbf{m}(\mathbf{B})$ for the MILES becomes expensive in terms of memory and computing time.

To overcome this issue, we implemented an efficient training where we randomly sample the training set of each fold so the optimization of $\mathbf{m}(\mathbf{B})$ is manageable, while maintaining samples from all subjects and enough information for classifier

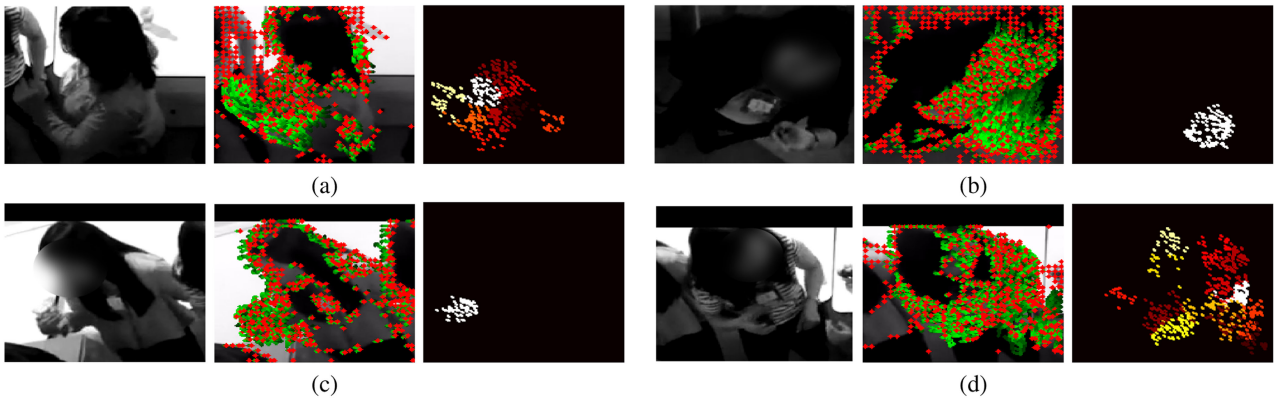


Fig. 6. Analysis of instances with higher contribution in the MILES classifier. From left to right: Original image, original dense trajectories, trajectories considered as of high contribution. Different colors in right images represent different clusters (see Section IV-A). (a) MILES focusing on trajectories corresponding to the hand/arm area (b) Background trajectories (e.g. shadows) beign omitted. (c) MILES handling cross-contamination of subjects. (d) Failure case, MILES also considers gesture from another participant.

regularization. To find the optimal value for this trade-off (enough samples versus memory limitations) we train with a maximum number of bags ranging from 100 to 5,000. This experiment showed that at 1,500 bags the results start saturating and adding more samples has no evident benefit. Hence, we chose this value for all our following experiments. Also, our data has a strong imbalance between positive and negative bags for most subjects (positive $>>$ negative), resulting in a heavily imbalanced training set. To overcome this we do the sampling in a stratified manner.

In contrast, for the test set in each fold we used the entire set of bags $\mathbf{B}^s$ for the subject left out. We used the AUC as evaluation metric instead of the accuracy to account for the imbalance in our samples. For the classification we used the MILES implementation in PRTools [38]. Applying this methodology we obtained a mean AUC of 0.67 ± 0.09 with 68 subjects (2 had issues for the segment and had to be discarded).

In addition, using the same training samples given to the MILES, we trained a Fisher classifier which is used as a baseline comparison for the video. We can see in Table I that the use of MILES with the same set of samples further improves the classification, when compared to this baseline classifier.

1) Analysis of Bag Instances: As stated before, we are interested in analyzing qualitatively which instances in the bags contributed the most to the MILES classifier, to assess whether they correspond to trajectories of the correct subject. We do so by calculating the contribution of each instance in a new test bag, using Eq. 4 as described in Section IV-A4.

Figure 6 shows example cases of instance importance used by the MILES classifier. From left to right we have the original image, the original dense trajectories, and those trajectories considered by the MILES classifier to be of high contribution following Eq. 4. To be considered as 'high', the absolute value of the instance contribution had to be higher than 0.0001 (to filter noise). Different colors in the images showing trajectories with high levels of contribution (right images) represent different clusters. Thus, if a prototype trajectory (see Section V-B) was selected as having a high contribution, all the trajectories in that cluster are shown.

First, in Figure 6(a) we can see an example of MILES having high contribution levels for trajectories corresponding to the arm an hand region of the person while a gesture is performed. This shows that the MILES classifier is learning the correct representations. Figure 6(b) shows that background trajectories are ignored while trajectories of the subject's hands are important.

Figure 6(c) shows the case where the MILES approach handles cross-contamination of subjects, assigning high contributions to the trajectories belonging to the subject owning the bounding box and no contribution to the trajectories from to the person causing the cross-contamination.

Finally, Figure 6(d) shows a failure case where trajectories corresponding to another subject (cross-contamination) are given high contribution. We will discuss further about these qualitative results in Section VI.

C. Decision Fusion

As stated before, to leverage the complementarity of both modalities we performed a decision level fusion using the posterior probabilities of both unimodal classifiers. We chose decision fusion as the bags in MILES are embedded into a different space than the features from the acceleration, so an early fusion approach is not appropriate.

Similar to the experiments in the past sections, we applied a leave-one-subject-out cross-validation, this time using Fisher's linear classifier to avoid overfitting.

This experiment gave us a mean AUC of 0.69 ± 0.10. So, as we hypothesized, the use of the complementarity between modalities increases the performance of the detection while compared to the unimodal approaches (0.67 and 0.65 for video and acceleration, respectively).

D. Impact of Noisy Data in Video on Performance

As seen in the deviation, there is still a high variability between the participants' results. Thus, we now proceed to analyze the levels of noisiness of the data, due to the cross-contamination between subjects and their occlusions, as a possible cause. This will be done in a static (in time) and dynamic manner.

TABLE II
ANALYSIS OF THE IMPACT OF DATA COMPLEXITY IN GESTURE DETECTION. MEAN AUC (± DEVIATION) USING LEAVE-ONE-SUBJECT-OUT AND DIFFERENT SUBSETS FOR TRAINING AND TESTING

Train (sub)set	Test (sub)set	Mean AUC ± std		
		Video	Wearable	Fusion
Entire	Entire	0.67 ± 0.09	0.65 ± 0.08	0.69 ± 0.10
Entire	Clean	0.67 ± 0.10	0.63 ± 0.07	0.68 ± 0.11
Entire	Noisy	0.69 ± 0.08	0.66 ± 0.09	**0.71 ± 0.09**
Clean	Entire	0.68 ± 0.09	0.65 ± 0.09	0.66 ± 0.10
Clean	Clean	0.66 ± 0.10	0.67 ± 0.09	0.66 ± 0.12
Clean	Noisy	0.70 ± 0.08	0.63 ± 0.11	0.66 ± 0.07
Noisy	Entire	0.68 ± 0.09	0.66 ± 0.10	0.69 ± 0.10
Noisy	Clean	0.67 ± 0.10	0.66 ± 0.11	0.68 ± 0.11
Noisy	Noisy	0.69 ± 0.08	0.70 ± 0.07	**0.71 ± 0.09**

1) Static Analysis of Noisy Data in Video: We first aimed to analyze the impact of the noisy data using a static comparison. Thus, we separated our set of subjects in 2 subsets: 1) the *clean* and 2) the *noisy* subjects. To select the subjects belonging to the *clean* subset we computed the overlapping ratio on each frame for all subjects and then computed their mean over time. For the visibility constraint, we used the annotations given by the dataset for *out of view*. We also added as *out of view* those moments for which the subject is too close to the borders of the frame and is only partially visible. A subject is part of the *clean* subset if their mean overlapping ratio is lower than 0.4 and is visible for 60% of the time. On the contrary, will be part of the *noisy* subset.

Table II summarizes the results while training with one of the subsets (or the entire set) and testing with another, using a leave-one-subject-out cross-validation. When training for a given subset we only used those subjects within the subset; leaving out the subject only if this is part of the subset selected for testing. For this table, the first row corresponds to the results presented in the previous subsections.

First, we can see in these results that the fusion column has in most cases a higher performance than the video classification. The exception are the experiments trained with the clean subset which suggests that the information added by the wearable devices in this case is redundant as the MILES has learned clear examples of gestures, and the information from the wearable is redundant. Also, note that overall the classifier using the wearable acceleration information performs similarly to the video. We will discuss this findings in Section VI.

2) Dynamic Analysis of Noisy Data in Video: Our aim with such comparison is to determine if it exists a correlation between those moments with high occlusion between participants (*noisy states*) and the confidence of the MILES classifier.

Similar to the static analysis, for each participant we calculated the overlapping ratio at each frame. In addition, we computed the distance of each person's to the closest image border and normalize it (border $= 1$, center of image $= 0$). The sum of these two on each frame give us a *ratio of occlusion* for each participant in time. Note that this ratio is not normalized as you can have a person in the border and occluded, for which ratio of occlusion > 1.

Figure 7 shows the error analysis in time for a subject chosen at random. Aside from the ratio of occlusion, these plots show the confidence of each classifier and the error for the MILES and the decision classifier.

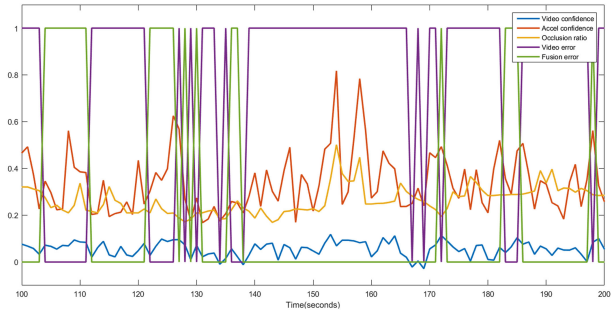


Fig. 7. Error analysis in time for a participant's errors and confidences.

TABLE III
SUMMARY OF SPEAK DETECTION RESULTS USING UNIMODAL CLASSIFIERS AND THEIR FUSION IN A DECISION-LEVEL. MEAN AUC (± DEVIATION) OF FOLDS

Classifier	Wearable Device	Video (Baseline)	Video (MILES)	Fusion
AUC	0.68 ± 0.09	0.54 ± 0.06	0.64 ± 0.11	0.64 ± 0.11

E. Detection of Speaking Status

Previous work in social psychology has found that there is a relation between the people's gestures and their speech [1], [2], [39]. Furthermore, works on automatic computing have leveraged this relationship in order to detect speaking status from the movement and/or gestures of the people while they interact [35], [37], [40].

Following the same premise, in this section we use our MIL method to detect binary speaking status (speaking/non-speaking) from the movement trajectories from video. In addition, we also applied our fusion approach with wearable information, and compare the results with those obtained for gesture detection.

Although MatchNMingle has groups interacting freely [28], we treat each participant independently when detecting its speaking status in a binary manner (speaking or not speaking). This means that we assumed that only one person is expected to speak at a time in our signals. Hence, as our method does not accounts for overlapping behavior during group interactions, more complex speaking concepts (such as speaker diarization [41]) are out of the scope of this paper.

To detect speaking status, we use the labels for the speaking status of all participants provided by the matchNMingle dataset. Speech, as well as the hand gesture gestures, was annotated every frame at 20 FPS. Table III summarizes the results for the unimodal and multimodal approaches. All these values are also statistically significant, with $p < 0.01$ when compared to a classifier assigning labels at random.

We can see in the results of Table III that, unlike gesture detection, the classification using the information from the wearable devices has the best performance. Moreover, this performance is similar to what was found in previous work using the same dataset [35]. In contrast, the results for the video (both baseline and MILES) are considerably lower than those found for gesture detection, and the fusion does not improve over the unimodal approaches. We will discuss more about these differences in Section VI.

VI. Discussion

Instance contribution to the MILES

We can see in Figure 6 that most of the trajectories chosen by the MILES as of high contribution are those corresponding to hands or arms. This applies even if other regions of the body for the subject, other subjects or the background are also moving (see Figure 6(b-c)). Thus, our MILES approach achieves, up to a certain level, its goal to compensate for cross-contamination in video.

Nonetheless, we also see in Figure 6(d) a failure case where trajectories corresponding to another subject (cross-contamination) are given high contribution. This case in particular was interesting, as during those segments the two subjects were engaged in a conversation and the subject causing the cross-contamination was also gesturing. Thus, the MILES learns correctly that these trajectories were corresponding to a gesture but fails to discriminate the subject.

Video occlusion and complementary modalities

We could see from the experiments in Section V-D that occlusion in video tends to affect the performance of the MILES classification, while the fusion of modalities compensate for this occlusion. First, in Figure 7 we can see how for those time with high occlusion ratio the MILES classifier makes error while the decision fusion compensates for these. In these intervals, although MILES is having trouble with confidence due to the noisy state, the wearable acceleration based classifier maintains or increases its confidence and allows the fusion classifier to correctly classify such moments.

This complementarity between the modalities can also be seen in Table II. First, we can see in these results that the fusion column has in most cases a higher performance than the video classification. The exception are the experiments trained with the clean subset which suggests that the information added by the wearable devices in this case is redundant as the MILES has learned clear examples of gestures, and the information from the wearable is redundant. This might also be the reason of the performance for the training with the clean subset and the test with the noisy one.

Note also that overall the classifier using the wearable acceleration information performs similarly to the video. And these values are also below the fusion except for the clean subset as training. This suggests that the wearable devices are not affected by the vision noisy states, and instead different factors cause the differences between the subsets. We hypothesize that these difference might be due to interpersonal differences, as suggested by Gedik *et al.* [35].

Gesture versus speaker status detection

Finally, we discuss the differences between the results find for gesture detection (Table I) and the speak status detection (Table III). This experiment tries to leverage the natural relationship between speech and gestures during conversations [1], [2].

Results show that for the speak status detection, the wearable device is more informative than video, and the values were similar to those found in previous work using the same dataset [35]. Nonetheless, in contrast with the gesture detection, the results for the video (both baseline and MILES) are considerably lower than those found for gesture detection, and the fusion does not improve over the unimodal approaches.

The frequency distribution of each action could help understanding why our method is only partially working for the speaking status, when compared to the gesture detection.

We found that on average participants spent 407 seconds speaking (deviation per participant of 259 seconds) and 308 gesturing (deviation of 200 seconds). Nevertheless, the two actions overlap for only 196 seconds (deviation of 168). This analysis shows that i) not all the gestures are related to speech and ii) not all speech was strictly accompanied by a gesture. Thus, our method could try to interpret a gesture always as part of the conversation, which is not the case. Furthermore, as speech is not necessarily accompanied by a gesture, we might have a high number of false negatives for moments with only speech. This could also explain why the wearable is more accurate on detecting the speaking status than the video or fusion, as the devices sense movement mainly from the torso, which is more likely to move when a person speaks, and not necessarily the hands/arms.

VII. Conclusions

In this work we presented our method to detect gestures during crowded mingle scenarios using bags of dense trajectories from video and wearable acceleration. This detection is particularly complex for mingle scenarios as they present high subject cross-contamination and strong occlusions, among other additional challenges.

To overcome the highly noisy video data we applied a multiple instance learning approach (MILES), which showed to be able to handle problems such as non-static backgrounds and the cross-contamination between subjects up until certain point. Also, we analyzed the contribution of the instances in the classifier showing that this learns from trajectories representing a person gesturing and ignores those from the background, for example.

Leveraging the decision fusion of the video and wearable modalities shows an improvement in the detection performance, with a mean AUC of 0.69 ± 0.10 (compared to 0.67 and 0.65 for video and acceleration, respectively).

We also investigated the impact on the performance of noisy data due to subject cross-contamination and occlusions, both in a static and dynamic (in time) manner. This analysis showed that fusing modalities also compensates for those moments where the confidence of the MILES classifiers decays, due to occlussions. Finally, we applied our method to detect binary speaking status, leveraging the premise that gestures and speak are generally intertwined.

References

[1] A. Kendon, *Gesture: Visible Action as Utterance*. Cambridge, U.K.: Cambridge Univ. Press, 2015.

[2] D. McNeill, *Hand and Mind: What Gestures Reveal About Thought*. Chicago, IL, USA: Univ. Chicago Press, 1992.

[3] R. M. Krauss, Y. Chen, and P. Chawla, "Nonverbal behavior and nonverbal communication: What do conversational hand gestures tell us?," *Adv. Exp. Social Psychol.*, 1996.

[4] S. Escalera *et al.*, "Multi-modal gesture recognition challenge 2013: Dataset and results," in *Proc. Int. Conf. Multimodal Interact.*, 2013, pp. 445–452.

[5] J. Wan *et al.*, "ChaLearn looking at people RGB-D isolated and continuous datasets for gesture recognition," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. Workshops*, 2016, pp. 761–769.
[6] P. Wang, W. Li, Z. Gao, C. Tang, and P. O. Obunbona, "Depth pooling based large-scale 3-d action recognition with convolutional neural networks," *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1051–1061, May 2018.
[7] M. Asadi-Aghbolaghi *et al.*, "A survey on deep learning based approaches for action and gesture recognition in image sequences," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, 2017, pp. 476–483.
[8] X. Alameda-Pineda *et al.*, "SALSA: A novel dataset for multimodal group behavior analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1707–1720, Aug. 2016.
[9] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image Vision Comput.*, vol. 27, no. 12, pp. 1743–1759, 2009.
[10] Z. Yang, A. Metallinou, and S. Narayanan, "Analysis and predictive modeling of body language behavior in dyadic interactions from multimodal interlocutor cues," *IEEE Trans. Multimedia*, vol. 16, no. 6, pp. 1766–1778, Oct. 2014.
[11] X. Alameda-Pineda, Y. Yan, E. Ricci, O. Lanz, and N. Sebe, "Analyzing free-standing conversational groups: A multimodal approach," in *Proc. ACM Int. Conf. Multimedia*, 2015, pp. 5–14.
[12] Y. Chen, J. Bi, and J. Z. Wang, "MILES: Multiple-instance learning via embedded instance selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 1931–1947, Dec. 2006.
[13] N. C. Camgoz, S. Hadfield, O. Koller, and R. Bowden, "Using convolutional 3d neural networks for user-independent continuous gesture recognition," in *Proc. 23rd Int. Conf. Pattern Recognit.*, 2016, pp. 49–54.
[14] P. Wang *et al.*, "Large-scale continuous gesture recognition using convolutional neural networks," in *Proc. 23rd Int. Conf. Pattern Recognit.*, 2016, pp. 13–18.
[15] X. Chai, Z. Liu, F. Yin, Z. Liu, and X. Chen, "Two streams recurrent neural networks for large-scale continuous gesture recognition," in *Proc. Int. Conf. Pattern Recognit.*, 2016, pp. 31–36.
[16] Z. Ren, J. Yuan, J. Meng, and Z. Zhang, "Robust part-based hand gesture recognition using kinect sensor," *IEEE Trans. Multimedia*, vol. 15, no. 5, pp. 1110–1120, Aug. 2013.
[17] C. Wang, Z. Liu, and S.-C. Chan, "Superpixel-based hand gesture recognition with kinect depth camera," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 29–39, Jan. 2015.
[18] H. Liang, J. Yuan, and D. Thalmann, "Parsing the hand in depth images," *IEEE Trans. Multimedia*, vol. 16, no. 5, pp. 1241–1253, Aug. 2014.
[19] J. Alon, V. Athitsos, Q. Yuan, and S. Sclaroff, "A unified framework for gesture recognition and spatiotemporal gesture segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 9, pp. 1685–1699, Sep. 2009.
[20] S. Rautaray and A. Agrawal, "Vision based hand gesture recognition for human computer interaction: A survey," *Artif. Intell. Rev.*, vol. 43, pp. 1–54, 2015.
[21] Y. Xiong and F. Quek, "Hand motion gesture frequency properties and multimodal discourse analysis," *Int. J. Comput. Vision*, vol. 69, pp. 353–371, 2006.
[22] F. Quek *et al.*, "Multimodal human discourse: Gesture and speech," *ACM Trans. Comput.-Human Interact.*, vol. 9, pp. 171–193, 2002.
[23] A. Marcos-Ramiro, D. Pizarro-Perez, M. Marron-Romera, L. Nguyen, and D. Gatica-Perez, "Body comunicative cue extraction for conversational analysis," in *Proc. IEEE 10th Int. Conf. Workshops Autom. Face Gesture Recognit.*, 2013, pp. 1–8.
[24] A. Marcos-Ramiro, D. P. Perez, M. Marron-Romera, and D. Gatica-Perez, "Capturing upper body motion in conversation: An appearance quasi-invariant approach," in *Proc. 16th Int. Conf. Multimodal Interact.*, 2014, pp. 327–334.
[25] A. Cerekovic, O. Aran, and D. Gatica-Perez, "Rapport with virtual agents: What do human social cues and personality explain?," *IEEE Trans. Affective Comput.*, vol. 8, no. 3, pp. 382–395, Jul.–Sep. 2017.
[26] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 2, pp. 288–303, Feb. 2010.
[27] Y. Yi and M. Li, "Human action recognition with graph-based multiple-instance learning," *Pattern Recognit.*, vo. 53, pp. 148–162, 2016.
[28] L. Cabrera-Quiros, A. Demetriou, E. Gedik, L. van der Meij, and H. Hung, "The matchnmingle dataset: A novel multi-sensor resource for the analysis of social interactions and group dynamics in-the-wild during free-standing conversations and speed dates," *IEEE Trans. Affective Comput.*, to be published.
[29] P. Atrey, M. A. Hossain, A. E. Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: A survey," *Multimedia Syst.*, vol. 16, pp. 345–379, 2010.
[30] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," Colorado Springs, CO, USA, pp. 3169–3176, Jun. 2011. [Online]. Available: http://hal.inria.fr/inria-00583818/en
[31] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *Int. J. Comput. Vision*, vol. 103, pp. 60–79, 2013.
[32] J. C. van Gemert, M. Jain, E. Gati, and C. Snoek, "APT: Action localization proposals from dense trajectories," in *Proc. Brit. Mach. Vision Conf.*, 2015, pp. 177-1–177-12.
[33] R. Stewart, M. Andriluka, and A. Y. Ng, "End-to-end people detection in crowded scenes," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 2325–2333.
[34] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2016, pp. 779–788.
[35] E. Gedik and H. Hung, "Personalised models for speech detection from body movements using transductive parameter transfer," *Pers. Ubiquitous Comput.*, vol. 21, pp. 723–737, 2017.
[36] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Adv. Large Margin Classifiers*, 1999, pp. 61–74.
[37] H. Hung and E. G., and L. Cabrera-Quiros, "Detecting conversing groups with a single worn accelerometer," in *Proc. ACM Int. Conf. Multimodal Interact.*, 2014, pp. 84–91.
[38] P. Duin *et al.*, "Prtools4. 1, a matlab toolbox for pattern recognition." Delft University of technology, vol. 2600, 2007.
[39] A. Kendon, *Conducting Interaction: Patterns of Behavior in Focused Encounters*. Cambridge, U.K.: Cambridge Univ. Press, 1990.
[40] M. Cristani, A. Pesarin, A. Vinciarelli, M. Crocco, and V. Murino, "Look at who's talking: Voice activity detection by automated gesture analysis," *Int. Joint Conf. Ambient Intell.*, vol. 277, pp. 72-80, 2011.
[41] X. Anguera *et al.*, "Speaker diarization: A review of recent research," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 356–370, Feb. 2012.

Laura Cabrera-Quiros received the "Licenciatura" and M.Sc. degrees from the Instituto Tecnólogico de Costa Rica, in 2012 and 2014, respectively, and the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands, in 2018. She is a Guest Postdoctoral Researcher with the Socially Perceptive Computing Group, Delft University of Technology. She is also a Postdoctoral Researcher with the Eindhoven University of Technology, Eindhoven, The Netherlands and with the Maxima Medical Center, Eindhoven, The Netherlands. Her main interests are the use and fusion of wearable sensing and computer vision for applications oriented to the analysis of human social behavior and health monitoring.

David M. J. Tax received the M.Sc. degree in physics from the Radboud University Nijmegen, Nijmegen, The Netherlands, in 1996, and the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands, in 2001. He was a Marie Curie Fellow with the Intelligent Data Analysis Group, Berlin, Germany. He is currently an Assistant Professor with the Pattern Recognition Laboratory, Delft University of Technology. His current research interests include the learning and development of detection algorithms and (one-class) classifiers that optimize alternative performance criteria, and multiple instance learning.

Hayley Hung received the Ph.D. degree in computer vision from Queen Mary University of London, London, U.K., in 2007, and her first degree in electrical and electronic engineering from the Imperial College London, London, U.K. She is an Associate Professor Head with the Socially Perceptive Computing Group, Delft University of Technology, Delft, The Netherlands. Between 2010 and 2013, she held a Marie Curie Fellowship with the Intelligent Systems Lab, University of Amsterdam, Amsterdam, The Netherlands. Between 2007 and 2010, she was a Postdoctoral Researcher with the Idiap Research Institute, Martigny, Switzerland. Her interests are social computing, social signal processing, computer vision, and machine learning.