A Hierarchical Approach for Associating Body-Worn Sensors to Video Regions in Crowded Mingling Scenarios

Cabrera Quiros, Laura; Hung, Hayley

DOI
10.1109/TMM.2018.2888798

Publication date
2019

Document Version
Final published version

Published in
IEEE Transactions on Multimedia

A Hierarchical Approach for Associating Body-Worn Sensors to Video Regions in Crowded Mingling Scenarios

Laura Cabrera-Quiros and Hayley Hung, *Member, IEEE*

***Abstract*—We address the complex problem of associating several wearable devices with the spatio-temporal region of their wearers in video during crowded mingling events using only acceleration and proximity. This is a particularly important first step for multisensor behavior analysis using video and wearable technologies, where the privacy of the participants must be maintained. Most state-of-the-art works using these two modalities perform their association manually, which becomes practically unfeasible as the number of people in the scene increases. We proposed an automatic association method based on a hierarchical linear assignment optimization, which exploits the spatial context of the scene. Moreover, we present extensive experiments on matching from 2 to more than 69 acceleration and video streams, showing significant improvements over a random baseline in a real-world crowded mingling scenario. We also show the effectiveness of our method for incomplete or missing streams (up to a certain limit) and analyze the tradeoff between length of the streams and number of participants. Finally, we provide an analysis of failure cases, showing that deep understanding of the social actions within the context of the event is necessary to further improve performance on this intriguing task.**

***Index Terms*—Mingling, wearable sensor, acceleration, computer vision, association.**

I. Introduction

SOCIAL gatherings such as parties, drinks receptions or networking events, provide an interesting study case to analyze group dynamics. Due to their ecological validity and social context these scenarios, which are commonly referred to as *mingling events*, have received increasing interest in recent years from the multimedia, computer vision, and ubiquitous computing communities [1]–[5], as example cases for automated human behavior analysis.

Manuscript received March 6, 2018; revised July 9, 2018 and October 10, 2018; accepted December 8, 2018. Date of publication December 20, 2018; date of current version June 21, 2019. This work was partially supported by the Instituto Tecnológico de Costa Rica. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Honggang Wang. *(Corresponding author: Laura Cabrera-Quiros.)*

L. Cabrera-Quiros is with the Department of Intelligent Systems, 2628 CD Delft, TU Delft, The Netherlands, and also with the Escuela de Ingeniería Electrónica, Instituto Tecnológico de Costa Rica, Cartago 30101, Costa Rica (e-mail: l.c.cabreraquiros@tudelft.nl).

H. Hung is with the Department of Intelligent Systems, TU Delft, 2628 CD Delft, The Netherlands (e-mail: h.hung@tudelft.nl).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author.

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.2018.2888798

Previous works in human behavior analysis have focused on smaller gatherings of people, such as meetings [6]–[10]. But unlike small group meetings, mingle scenarios have a higher number of people involved and a high number of social interactions dynamically occurring at the same time, which makes them more challenging for fined-grained group behavior analysis. In addition, people are not restricted to a predefined position or to follow a task or script, but rather can move and change conversational groups following their own affects.

Past efforts in human behavior analysis have proved that fusing modalities (e.g. video and audio or video and wearable sensors) increases the performance of recognition and classification of a wide variety of tasks, such as dominance [8], leadership [7] or cohesion [10]. Thus, each modality contributes to a different element of the event and acts as a complementary source of information. In addition, the use of multiple modalities had shown to be a suitable alternative to deal with challenging scenarios, including group gatherings [11], [12]. For instance, Alameda-Pineda *et al.* [5] showed improvements in the detection of head and body orientation and, consequently, in the analysis of free-standing conversational groups by leveraging the use of video, audio and wearable IR (Infra-Red).

Wearable devices are a modality that has been used considerably in mingle scenarios, due to its versatility. Works using this modality have presented encouraging results in human behavior understanding by analyzing people's movement as they interact [4], [13]. Thus, by leveraging video and wearable sensors one could analyze human behavior within these environments without interfering with the natural behavior of the people involved.

Nonetheless, although the use of wearable sensors as a complementary source of information has many advantages, manually associating a specific device to a particular region of the video (corresponding to the person using the device) quickly becomes a challenging practical issue as the number of streams to associate per modality increases, making the correct associations harder to discriminate.

In fact, when using other modalities along with video, the majority of works either i) manually associate video to the other modalities [5], [13], or ii) avoid the problem entirely by using only one source in the other modalities (e.g. only one wearable device or microphone) [7], [12].

In this paper, we present a solution to tackle this problem by associating the time series signals from wearable accelerometers to the acceleration streams extracted from video flows. Thus, we

aim to associate each device with the spatio-temporal region of video of its wearer. This association is particularly challenging for mingle scenarios, as people's social behavior in these events (unlike simple actions like walking or running) do not tend to have a predictable and easily distinguishable pattern; and as the number of people increases.

A preliminary version of this work was presented in [14]. There, we introduced a more simple version of our method to hierarchically associate wearable devices to the spatio-temporal region in video of the person wearing the corresponding device. To do so, we leveraged the use of proximity information obtained from the wearable devices and video as a spatial prior to the association process. Thus, we could applied a *divide and conquer* strategy, by associating the streams within all possible group combinations from different modalities, and then selecting the optimal group-to-group association (see Section IV-C3 for an overview of our method and its changes between works).

In this previous version, for groups with unequal instances of the modalities (e.g. more devices than people in video) the streams remaining after the group-to-group association were discarded. Also, the method could not handle missing data (e.g. person leaving the field of view of the cameras and then returning). In addition, all experiments were done with a limited number of participants (19). All the above resulted in a rather limited evaluation.

In contrast, in this paper we improve over the aforementioned aspects and present several novel contributions:

- We modify our method to account for unequal groups of streams and streams with missing data. Thus, our method is now optimized to handle any combination of streams, dynamically accounting for uneven numbers of streams both globally and in the group-to-group associations.
- We increase the number of streams to be associated to 69 participants in each modality. In addition, these streams have also missing data in different proportions (given the behavior of each person), which makes them a more suitable example of cases in real scenarios.
- We include a more comprehensive evaluation of our method. We address issues related with understanding the association process such as evaluating the impact of the number of participants and the period over which observations are accumulated on the accuracy of the association, the effects and errors introduced by the group-to-group matching and assess the impact on the association performance of missing streams, either partially (missing data) or completely (missing streams).
- We further analyze qualitatively if shared social actions (e.g. shared gestures or laughter) have any impact on the association process, as we hypothesized that due to mimicry these could become failure cases for our method.

The rest of the paper is arranged as follows. The work related to our own is described in Section II. The data used for our experiments is presented in III and our approach is described in Section IV. Section V presents the experimental procedure and Section VI shows our results. Section VII their discussion. Our conclusions are summarized in Section VIII.

II. Related Work

Several works have used information from video and wearable sensors for a wide range of tasks such as human action/activity classification [11], [12] or group detection [5], among many others. However, very few have addressed the challenging task of automatically associating the video pixels or regions with the additional sensor modalities, such wearable sensors. Although many works exists on video-to-audio association [15]–[17], which is generally called *speaker diarization*, we will only refer to works about association of video with wearable devices, as other modalities are outside the scope of this paper. For more details in audio-video association, see [17].

When associating wearable acceleration with video, previous works can be divided in 2 main approaches: 1) pixels-to-device association and 2) region-to-device association. In the former, each device is associated to the set of more similar pixels in terms of a given similarity measurement (e.g. Euclidean distance). As they have several more streams in one modality, such as the pixel trajectories of all people moving, these approaches tend to use 3D and orientation measurements in both modalities, which allows them to be more discriminative. In a region-to-device association, the set of pixels is previously clustered by a defined technique such as manual annotation, image-based segmentation or object tracking, among others. Then, each region of interest is associated with their corresponding device. Our work is an example of the latter.

Rofouei *et al.* [18] and Wilson and Benko [19] are examples of works using a pixels-to-device association. They proposed similar methods to match the 3D acceleration of a smartphone (also using its gyroscope) to the set of pixels with the higher similarity in a video recorded with a Kinect, which also recorded depth information. Thus, constructing the real 3D world coordinates from the Kinect and knowing its position w.r.t the real world, they mapped all the devices to these real world coordinates. To measure the similarity, their methods are based on an euclidean distance minimization between both streams. Bahle *et al.* [11] proposed a similar association of pixels-to-device, but limiting the pixels to those regions on the joints detected by the Kinect. They also used a 3D reconstruction of the real world and the Dynamic Time Warping (DTW) distance as similarity measure.

Although these methods essentially match acceleration streams like ours, their solutions are oriented to the interaction with a display using mobile phones. Hence, they do not consider a high number on devices and the implications that this could have in the association process with video. In addition, they reported problems with fast movements and during moments when the device was not moving.

For region-to-device association, the closest works to our own is Texeira *et al.* [20]. They presented an approach based on Hidden Markov Models to identify and localize moving smart phones (by their accelerometers and magnetometers) in a camera network. To do so, they modeled the association as a missing data problem where a person's behavior is observed twice, once by the camera and once by the wearable device, but the link between the 2 modalities is unknown. They proposed a solution that could ultimately work for more than one device, but in their

experiments have one single person walking under the network of cameras. This unique stream is later divide into 5 and each is treated as a different participant. This solution seems to be a suitable option to 'generate' more participants, but they do not address the challenges of occlusion resulting from a crowded scene making their solution infeasible for mingling groups. In addition, unlike in our case, the streams that they generate do not have any interaction between each other in real life, which makes the dataset they used not a nice representation of a mingle scenario, with its possible consequences in the matching process.

Other works in region-to-device association include Shigeta *et al.* [21], Plotz *et al.* [22] and Nguyen *et al.* [23]. The methods proposed by Shigeta *et al.* and Plotz *et al.* first detected the moving areas in the video and associate these to a corresponding device within a set of 5 and 3 devices, respectively. As their acceleration signal are not synchronized in time, unlike our case, they used the peaks in the Normalized Cross Correlation (NCC) between the acceleration signal and the region in the video to detect the proper alignment between the signals. Thus, once the peaks are found they choose the matches between devices and video using a greedy assignment.

These methods are feasible for a small number of devices but when this number increases the discrimination between the streams is harder to perform in a greedy manner, as we will prove later in this work (see Section VI-B), and the NCC starts to fail while providing the correct alignment. Also, both methods, are limited to moving objects.

Compared to these works (including Texeira's), our approach proposes a considerable increase in the number of accelerometers to be associated, where we show improvements in performance over the state-of-the-art methods even when matching over 60 video and wearable acceleration streams using a hierarchical grouping approach.

To the best of our knowledge, we are the first to consider the association of video with multiple wearable devices in such large and crowded scenarios, considering miss streams and streams with incomplete data. In addition, we propose to solve the association problem in a much more challenging context where people's behavior can not be as easily characterized as simple actions like standing and walking and it is harder to discriminate between people's movements.

III. Data

A. Real Mingling Scenario Dataset

For our experiments, we use the *MatchNMingle* dataset [24], where it was collected video and wearable acceleration for 92 participants during 3 separated group gatherings. This data was collected in a real mingling scenario after a speed date event, where people were encouraged to socialize. Due to hardware malfunctioning, 23 of the devices did not record data during the event leaving 69 functioning devices. Each person wore a custom-made wearable device, as the one seen in Fig. 1(a), hung around the neck which recorded triaxial acceleration at 20 Hz. This wearing method emulates a smart badge, making it feasible for speed dates, conferences or other type mingling social events. These devices also have a binary proximity

(a) (b)

Fig. 1. (a) One of our custom-made wearable devices (smart badge). (b) Snapshot of one camera from our mingle event.

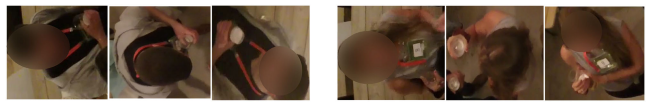

Fig. 2. Changes in appearance of 2 of our participants through the mingle event.

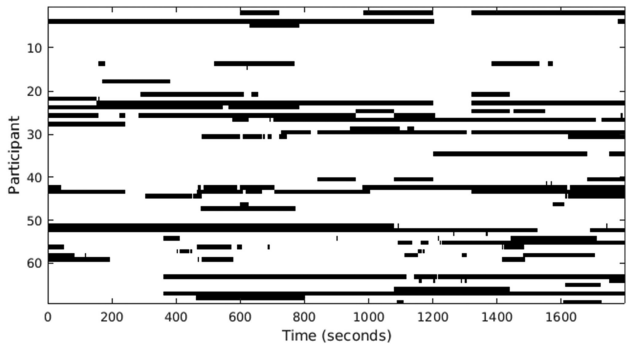

Fig. 3. Participants visibility status. Black = participant missing.

detector based on beacon communication with other devices. Thus, each device emits its own ID to all other devices around it. This beacon broadcasting allows the devices to synchronize every second and detect each other from 2-3 meters away. The detection of a device is considered as a proximity detection (binary signal). Overhead video was captured using 5 different GoPro Hero 3+ cameras that covered the whole mingling area with some overlap. A snapshots of our mingle event (from one of the cameras) is shown in Fig. 1(b).

Finally, 8 different social actions (Walking, Stepping, Drinking, Speaking, Hand Gesture, Head Gesture, Laugh and Hair Touching) were annotated for each participant (when visible) every second from video using the Vatic annotation tool [25]. These annotations were done by 6 different annotators, for which 2 minute intervals from the videos were given at random.

1) Complexity of Our Dataset: Since our event was recorded during a real mingle event, all the participants had the liberty to move around and leave the mingle area at will. They were recorded by different cameras, with different light conditions and strong appearance changes given their position w.r.t. the camera. For example, Fig. 2 shows the changes in appearance of 2 of our participants.

Moreover, Fig. 3 shows the visibility status of the 69 participants, under the 2 cameras with the higher concentration of people, for an interval of 10 minutes chosen randomly from the

TABLE I
NUMBER OF PARTICIPANTS VISIBLE FOR AT LEAST AN X AMOUNT OF TIME

Minimum time X under FoV (minutes)	1	2	3	4	5	6	7	8	9	10
Number of participants	54	53	52	51	50	46	46	37	33	22

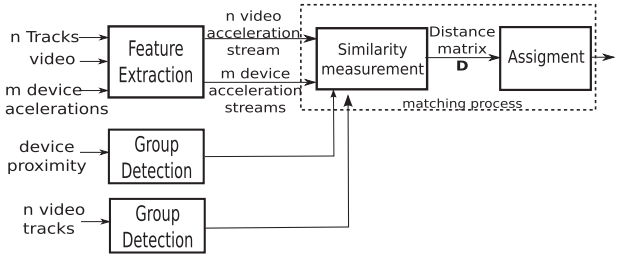

Fig. 4. Overview of our approach.

mingle segment. This is the same 10 minute interval that will be used later in our experiments. Notice that the visible times are not necessarily consecutive and that, in fact, some were missing for the entire interval.

From the entire 69 participants, 22 are visible for the entire 10 minutes while there is not video data for 14 people. The rest of participants are visible for a variable proportion of the time during the 10 minutes. Table I summarizes the number of participants visible for at least a given amount of time (X). Here, the subset were all people are under the FoV for the entire time (last column) will be our *ideal subset*, while the set with all 69 participants is our *entire set*. Thus, only a 31.9% of the streams are complete for our entire set while 20% of the streams are entirely missing.

IV. OUR APPROACH

Our approach is summarized in Fig. 4 and detailed below.

A. Feature Extraction

1) Wearable Devices: For the wearable devices, a single acceleration stream for each device is obtained using the magnitude of the 3 axes. Using magnitudes, instead of the 3 axes separated, allow us to compare the device's acceleration to the video without knowing the orientation reference between the two modalities in the real world. To eliminate the influence of gravity, each axis is first normalized using its mean and variance over the entire observation time.

2) Video: Each device stream must be assigned to a specific person in the video. As stated before, in this work we do not intend to perform a pixel-to-device association, but rather associate each device with a region containing a person. Hence, all those regions of interest (or bounding boxes), which include a person with a device, are first extracted. Then, we concatenate the bounding boxes over time for each person, to generate a track or tube (area of interest over time) for that person (see Fig. 5).

The Vatic tool [25] for video annotation was used to extract the bounding boxes. While this is a manual labeling tool, we found that using the SPOT tracker [26] gave us similar results, with a mean overlapping ratio between participants of 0.9006

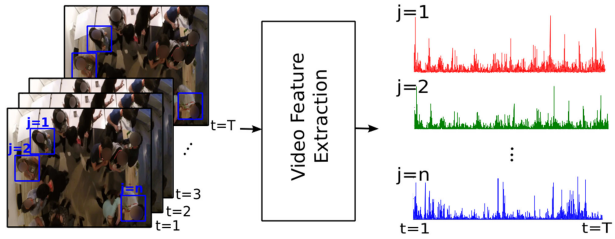

Fig. 5. Feature extraction from video for 3 example tracks (subjects). Output: speed stream for each participant for interval of length T.

(1 equals the highest) and a deviation of 0.0632 for 10 participants randomly selected in the 10 minute segment. However, as with all trackers, SPOT has a level of tracking noise. While heavy occlusion, tracking drifts and false detections are relevant problems in the tracking domain, this work focuses on the already challenging problem of associating large numbers of streams with relatively short observation times and not in tracking. We leave the investigation of this phenomena in relation to the association problem to future studies.

Once the areas are annotated and concatenated to form the position tubes for each person, we proceed to treat each tube as follows. First, we extract dense optical flow for the entire video. Then for a given bounding box, which belongs to a tube, we take the magnitude of all the flow vectors and then compute the mean for those with a magnitude greater than zero. In this way, we obtain a vector of mean flow magnitudes for a given tube over the entire video of length T (where T is the number of frames in the video interval). This is used to represent the velocity of movement for that person between two consecutive frames. This approach allows us to consider the influence of fine grained movements such as gestures or laughter as well as movement of the entire body. Fig. 5 shows a graphical representation of this process. Finally, we compute the acceleration vector from the speed using finite difference approximation to obtain a measurement comparable to the acceleration in the devices.

After we extract the acceleration streams from the video and wearable accelerometers, we proceed to treat each stream (video and device) as follows. First, we normalize the maximum value of all streams to one, so a comparison between video and wearable acceleration can be made. Next, we apply a sliding window calculating the variance over each stream. Using this instead of the raw acceleration will give us a better representation of the activity levels of the people [27]. Additionally, it has been proved in activity recognition using wearable acceleration that working with raw acceleration values can present difficulties due to recording noise, among others factors [28].

B. Similarity Metrics

Both video and acceleration streams are noisy because they capture only partially the behavior of a person. Since the device is hung around the neck, movements from the torso are strongly captured by this modalitiy. However, energetic gesturing in the video will not necessarily be directly translated into similarly energetic movement in the body. Therefore, we need measurements to assess how similar 2 streams are and not if they are equal. Different metrics are compared to quantify the affinity

between the acceleration streams from video and the devices: covariance (COV), Dynamic Time Warping Distance (DTW) and Mutual Information (MI). These metrics are widely used to assess affinity between streams [29].

Notice that in our previous work [14], the similarity metrics needed complete streams. Here, we have improved our method, which can now handle streams with missing data. Now, the similarity metrics only consider those sections where there is information for both modalities. Hence, for the covariance and the mutual information, intervals with missing data in one or both modalities are ignored, and weighted given the length of the complete stream. For the Dynamic Time Warping distance, all sections of complete data are treated as separate streams and the overall distance is calculated by taking the mean distance of all segments weighted by their length.

C. Assignment Methods

We consider the matching process to be an assignment problem, where m elements of a set M (device streams, in our case) need to be associated with n elements of a set N (video streams), by fulfilling a given function or constraint. The distances matrix $\mathbf{D}_{ij}$, of size $m \times n$, is formed by the pairwise distances between all possible combinations of m acceleration and n video streams, where

$$\mathbf{D}_{ij} = d(i,j), i \in \{1 \cdots m\} \text{ and } j \in \{1 \cdots n\} \tag{1}$$

and d is one of the similarity metrics in Section IV-B.

1) Winner-Takes-All (Greedy) Association: State-of-the-art methods ([18], [19], [21]) use a greedy approach where the element in $\mathbf{D}_{ij}$ that has the highest value determines the assignment. The corresponding column and row are removed from $\mathbf{D}_{ij}$ and the assignment process is repeated. This relies on a strong correlation between the sensor data for a given device and its corresponding video stream. This is the baseline that we compare our proposed method with.

2) Hungarian Method: Although the winner-takes-all method is a reasonable baseline, it does not consider that there is likely to be noise in both sensor streams. Hence, it may not be able to distinguish one possible assignment from the other. This is particularly problematic as the number of streams increases. In this case, trying to optimize the assignments globally may help.

The Hungarian method [30] computes a solution for the linear assignment problem by optimally matching the elements m and n, based on a global optimization of $\mathbf{D}_{ij}$. For this assignment problem, given the matrix of distances $\mathbf{D}_{ij}$, the aim is to find the global cost c that minimizes

$$\begin{aligned} \min & \sum_{i=1}^{m}\sum_{j=1}^{n} d(i,j) * w(i,j) \\ \text{s.t.} & \sum_{i=1}^{m} w(i,j) = 1,\ j = 1,2,\ldots,n \\ & \sum_{j=1}^{n} w(i,j) = 1,\ i = 1,2,\ldots,m \\ & w(i,j) = 0,1 \end{aligned} \tag{2}$$

where $w(i,j)$ is the binary weight for matrix $\mathbf{W} \in \{0,1\}^{m \times n}$ for the element (i,j). Thus, $w(i,j) = 1$ if the two pairs are associated, the method will choose the pairs of elements with the lowest total pairing cost and the elements of sets M and N can only be paired once. Several solutions exist to solve this problem [31]. Notice that Eq. (2) is defined such that it holds for those cases where $m \neq n$. Thus, our association is not limited to an equal number of streams on each modality.

3) Hierarchical Hungarian Method: As the number of streams to be associated increases there is a higher probability of finding 2 or more people with similar streams. Hence, the observation period needs to be longer to increase the chances of discriminating between them. Computationally speaking, however, it is desirable for a potential real-time application of this work, to be able to rely on shorter time intervals to make the association. Although, if the streams are too short, we will not have enough observable behavior for the distance metric to be discriminative enough.

By initially sub-dividing the problem based on the local spatial neighborhood in each sensor, we hypothetically could improve the numbers of correctly associated streams. Therefore, we propose an extension to the original Hungarian method by performing the assignment procedure in a hierarchical manner using a divide-and-conquer strategy where all the streams are subdivided into groups in each modality. This reduces the problem initially to a smaller size represented by the number of groups in each modality. We propose to generate the groups by clustering based on their proximity over a particular time interval (described later in Section IV-D). This further reduces the assignment problem from a global to local assignment problem, which exploits the local-spatial and social context of the mingling gathering.

So, for this new assignment methods, the n video and m accelerometer streams are clustered into p groups for the acceleration and q groups for the video streams. Then, $p \times q$ different distance matrices are generated; one for each possible group combination (e,f) where indices $e \in \{1 \cdots p\}$ and $f \in \{1 \cdots q\}$. For each of these matrices, the corresponding stream assignment is calculated. So, within each group-to-group matching, the possible stream combinations are now reduced to $n'_e \times m'_f$, where n'_e and m'_f are the number of elements in the eth and fth device and video groupings, respectively.

The cost $c(e,f)$ of each group-to-group assignment is then obtained by Eq. (2). These costs are allocated in a new matrix $\mathbf{C}$, which represents the costs of assigning the elements within each possible group combination e and f. Note that each cost $c(e,f)$ must be normalized by dividing by the number of total costs that were used in each assignment so $\mathbf{C}(e,f) = c(e,f)/\min(m'_f, n'_e)$. For example, when comparing a group of 3 streams against a group of 2, only 2 costs from the 3×2 matrix are used for the final assignment, whereas comparing 2 groups of 3 streams we will have a summation of 3 costs from the 3×3 matrix.

Finally, the Hungarian algorithm is applied to matrix $\mathbf{C}$ to find the optimal group-to-group assignment. The stream assignment for that specific group-to-group pairing is then chosen. An example of our Hierarchical Hungarian assignment procedure is illustrated in Fig. 6.

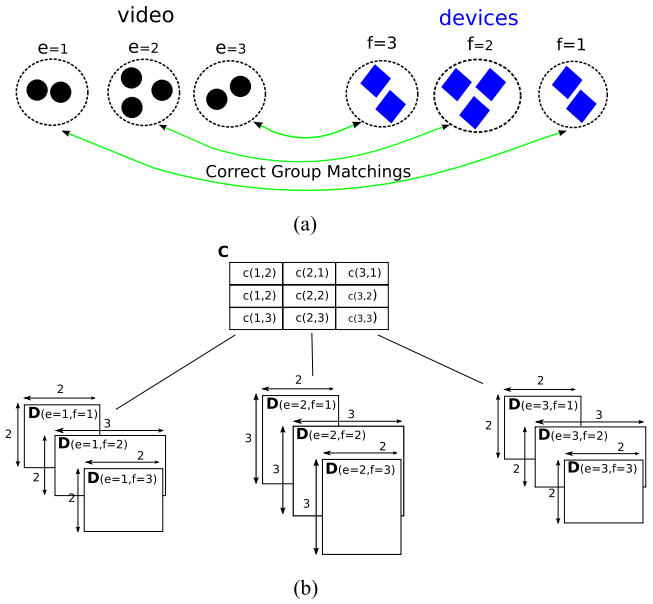

Fig. 6. Example of assignment method. (a) Devices and Video streams representations. The dotted circles show the group detection. (b) Our proposed *Hierarchical Hungarian* method using the streams and clusters from (a).

Notice that due to a mismatch in the number of streams in each modality (e.g. missing people or incorrectly matched groups), some streams can be left without associating. To account for these cases, we modify our original method (presented in [14]).

First, all similarity metrics used in the assignment already account for missing data, as explained in Section IV-B. Secondly, from our previous submission we noticed that when the groups are wrongly matched some streams are left unassigned even when one or more streams are left over in each modality. Thus, in this paper we improve our previous method and make it resilient against incorrectly matched groups or incorrect group detections in either modality, as the proximity prior is noisy and imperfect. To do so, when there is one or more streams in *both* modalities, we performed a final assignment with the remaining streams, without grouping, treating all the streams as singletons (group of one person). Those streams remaining without association after this final step are treated as missing people in one of the modalities ($m \neq n$ in Eq. (1)). Thus, the improved method in this paper can also handle uneven number of streams.

D. Group Detection

For group detection we use the method proposed by Hung and Kröse [1]. In this approach, the group detection is performed independently per sensor type using mainly the same clustering algorithm based in maximal cliques, treating the proximities of the participants as graphs. The difference between the 2 modalities lies in the process to create such graph. For each sensor type, this process and the clustering is described below.

The reason for using Hung and Kröse's approach [1] is that their method allows orientation to be implicitly represented based on the relative position of the people. This has a good analog to the wearable sensor data which, while being sensitive to detecting badges at certain orientations, also does not record precise orientation information either. In contrast, other methods for detecting groups ([3], [32]) generally required the orientation information explicitly, which we do not have. Also, [1] provides an accurate approximation for conversational groups, following the same scheme as the behavior presented by people during group forming.

1) Clustering Video Streams: To create the graph for cluster the video streams, we use the tracks extracted for each of the participants, focusing in the position of the center of the track in each frame of the video. Thus, for each frame, an affinity matrix $\mathbf{A}$ is created, which defines a symmetric distance between person i and j

$$\mathbf{A}_{ij} = -e^{\frac{d_{ij}}{2\sigma^2}} \tag{3}$$

where d_{ij} is the Euclidean distance in the image plane between the centroids of the bounding boxes for person i and j and σ is the width of the Gaussian kernel. In our experiments, σ was set to 150 pixels, as this was an approximate value for group distance given the image size and resolution of the camera. This threshold was selected by learning the mean area of coverage of all our participants in video for the entire dataset.

Then, we apply the group detection algorithm that extracts clusters as maximal cliques in edge-weighted graphs [1]. This is an iterative procedure that optimizes the group clustering based on the notion of a dominant set. If we have a graph G with each node representing the centroid of a person's bounding box and the affinity between people to be the edges, we can consider a representation of the closeness of a subset S of the graph as follows. We define a measure called the average weighted degree of a vertex $i \in S$ with respect to set S as $k_S(i) = \frac{1}{|S|}\sum_{j \in S} a_{ij}$. The relative affinity between node $j \notin S$ and i is defined as $\phi_S(i,j) = a_{ij} - k_S(i)$, and the weight of each i with respect to a set $S = R \cup \{i\}$ is defined recursively as

$$w_S(i) = \begin{cases} 1 & \text{if} |S| = 1 \\ \sum_{j \in R} \phi_R(j,i)\, w_R(j) & \text{otherwise} \end{cases} \tag{4}$$

$w_S(i)$ measures the overall relative affinity between i and the rest of the vertices in S, weighted by the overall affinity of the vertices in R. Therefore to find the cliques in the graph $w_S(i) > 0, \forall i \in S$. For every graph, only one maximal clique can be identified at a time and a peeling strategy is employed where the same conditions are repeatedly applied to the remaining sub-graph until no more cliques remain.

Finally, the cliques selected per frame are combined into a single set of groupings q for the entire video segment using majority voting. Thus, groups with the same set of participants are counted for the entire segment of recordings and the ones with the exclusive majority are selected.

Luckily, in a mingle scenario the people tend to stay in the same group for long intervals of time, making this selection method feasible. For example, for our event 17% of the participants stayed in the same group for the entire 10 minutes, 20% joined only 2 groups, 11% joined 3 and 17% joined 4 groups (total of 65%). Only 17% joined 6 or more groups. Notice that these statistics includes merging groups and excludes singletons

TABLE II
SIZE DISTRIBUTION OF THE GROUPS FOR OUR MINGLE SCENARIO

Group size	1	2	3	4	5+	Total
Numb. groups	15	15	7	2	1	40

(people alone). Thus, 2 groups of 2 people joining counts as 2 groups joined, even if they stayed with a same person during the entire event.

During the entire event (30 minutes), we had the total group size distribution presented in Table II. This distribution shows the different conditions in which our method would be able to perform regarding group sizes. Note that these groups overlapped, merged and split at different moments over time, and involved the same people in some cases. So, the sum of groups in Table II is not equal to the total number of people.

In practice, if the groups vary more frequently over time it would be rather straightforward to compute groups over short sliding windows, thus performing the association dynamically in time (see Section VI-B for an analysis on window lengths for our method).

2) Clustering Devices: As stated in Section III, each of the wearable devices outputs a dynamic binary proximity graph, which is later refined to eliminate false neighbor detections using the method proposed by Martella *et al.* [33]. Thus, for each time sample which is recorded at the same sample rate as the video (20 Hz against 20 fps) a proximity graph is created between the participants. To refined false neighbor detections, they apply a density-based clustering to group all the neighbor detections in time, by comparing the graphs of consecutive times. This method leverages the bursty nature of the proximity graphs, meaning that the correct neighbor detections tend to appear sequentially together in time and the false detections tend to be isolated (see [33] for more details).

Finally, the maximal cliques are identify from the proximity graphs, to obtain p sets of fully connected nodes, using the same maximal clique methods as with video. Here, d_{ij} in the affinity matrix $\mathbf{A}$ from Eq. (3) is created with the binary values from the proximity graphs.

V. EXPERIMENTAL PROCEDURE

A. Extraction of Acceleration Streams

For our experiments, we selected a 10 minute interval chosen randomly in the middle of the mingle event. For all 69 people with functioning devices, we extracted our wearable acceleration streams (see Section IV-A1) using a sliding window of 50 samples with a shift of one sample for which we calculated the variance. This window length (equals to 2.5 seconds) gives enough time for an human action to fully develop.

As explained in Section III, not all participants were present under the FoV of the cameras for the entire interval. So, the video acceleration streams were extracted for these 69 participants were video data was available. If their video data was incomplete, the acceleration stream was set to zero for those times only. This is done for purposes of a further comparison with our old method (see Section VI-D). Nonetheless, as explained in Section IV-B, these sections are not taken into account for the creation of our distance matrix with our new approach.

TABLE III
ASSOCIATION ACCURACY WITHOUT GROUPING FOR THE IDEAL SUBSET (22 PARTICIPANTS) AND THE ENTIRE SET (69 PARTICIPANTS)

	Accuracy (%)						
	Random	Greedy			Hungarian		
	baseline	MI	COV	DTW	MI	COV	DTW
Ideal	4.5	36.36	63.64	18.18	22.73	**81.82**	77.27
Entire	1.45	11.59	37.68	11.59	13.04	**46.38**	36.23

B. Accuracy Metrics

In general, we will treat as true positives (TP) all the pairs of streams that were associated correctly. Thus, our association accuracy will be number of true positives over the total number of streams to associate in the modality with less streams, or $acc = \frac{TP}{min(m,n)}$. Notice that, as well as Eq. (2), this considers a different number of streams on each modality. Also, in those cases with K-folds (e.g. leave out experiments), the mean accuracy will be the equal to acc_{fold}/K.

For the association including grouping (see Section IV-C3), the accuracy will be equal to the number of true positives that were correctly associate within a group matching that was also correctly associated. Also, TP_{group} will be used to denote those groups that were correctly matched and $acc_{group}(e, f)$ as the association accuracy within a given group pair (e, f).

VI. EXPERIMENTAL RESULTS

A. Comparing Between Distance Metrics (Without Grouping)

First, we compare the metrics in Section IV-B. Our intention is to assess the impact of each metric on the original linear assignment problem without applying our hierarchical approach just yet. To do so, we used our *ideal subset* (22 people, as seen Table I) where there is not missing data which represents an ideal scenario and our entire set of 69 participants. For both sets we used the entire segment of 10 minutes. Also, for the participants with missing video the acceleration streams from video were set to zero.

Table III summarizes the results for the association of both sets. For both sets, all similarity metrics (using greedy or Hungarian) outperform the random baseline. Using the covariance (COV) as a similarity metric gives the best association performance for either the greedy or the Hungarian assignment. The DTW seems to work well when combined with the Hungarian assignment, suggesting that this metric generates similarity values which are close together while the COV gives more discriminative values. Hence, a global optimization is necessary for the DTW but not for the COV. We will discuss more about the difference between the methods and metrics in Section VII.

We also found a significant difference in the association accuracy between our ideal subset (22 people with only clean data) and our entire set (69 people with missing data). This difference could be explained by one of 3 factors (or a combination of them): 1) different number of participants, 2) quality of the data

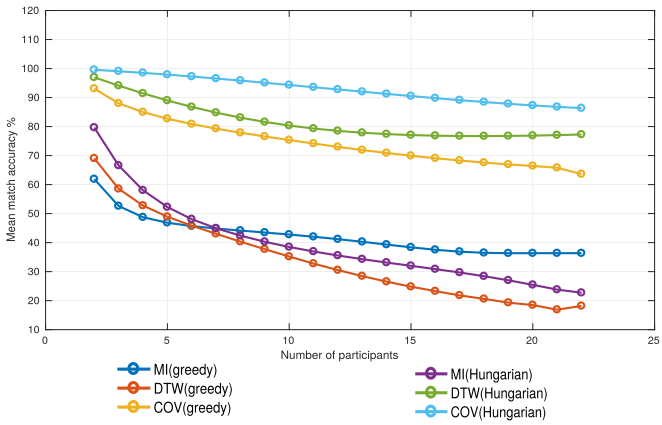

Fig. 7. Mean association accuracy without grouping for the different number of participants using the ideal set as base (10 minutes). Figure best seen in color.

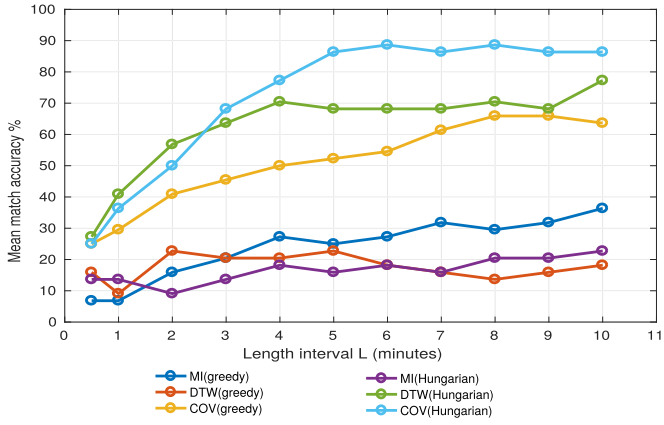

Fig. 8. Association accuracy without grouping for the different length intervals using the ideal set (22 participants). Figure best seen in color.

(clean versus missing) or 3) an social aspect within the observations. These 3 aspects will be further developed in the next subsections.

B. Effects of the Number of Streams and the Interval Length on the Association Process

In this section we analyze the impact on the association accuracy of the number of participants to be associated and the observation length when extracting the acceleration streams. To do so, we use only our *ideal set* as base to maintain clean conditions (e.g. no missing data).

First, for the analysis of the impact of the number of participants on the performance, we run associations with different number of participants. On each run, were $N \in \{2, \ldots, 22\}$ participants (ideal set has a total of 22 streams), we leave out k different participants iteratively ($k = 22 - N$) considering all possible K tuples. We then calculated the mean accuracy obtained by each association with N streams. Fig. 7 summarizes these results.

We can see how there is only an accuracy difference of about 13% between the sets of 22 and 2 participants, even when the number of participants was increased by a factor of 10. This was possibly due to the long interval (10 minutes) that was used for the association and supports our finding in [14] regarding the strong trade-off between streams to associate (participants) and observation time when the association is done without grouping. Given that shorter observation intervals are preferred and supported by the results in the last rows of the Fig. 7, we opted for a group-to-group assignment (Section VI-D).

Now, to analyze the impact of the length of the observation time, we gradually decrease this interval for the extraction the acceleration streams and calculated the association accuracy. Given that different parts of the interval can have different actions/events, we calculated the association using a sliding window of length L and shift it by $L/2$ and then report the mean value with its deviation over all the intervals. Fig. 8 shows these results. Here, at least an observation time of 5 minutes is needed to accurately associate more than 80% of the 22 streams. A similar result was found in [14].

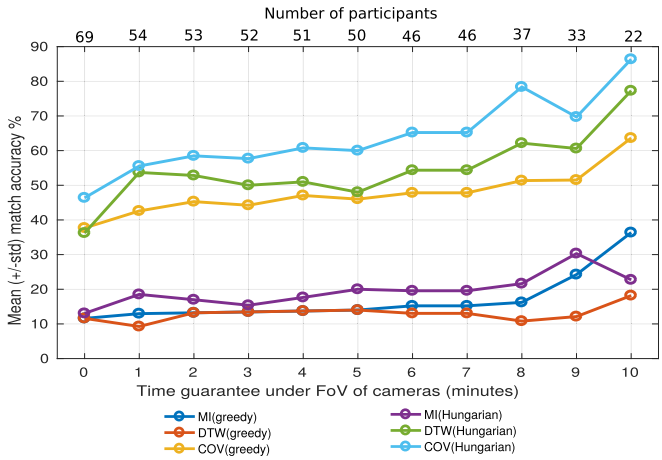

Fig. 9. Association accuracy without grouping for the different subsets of participants in Table I (10 minutes interval). Figure best seen in color.

C. Impact of Missing People in Video

As seen in Table I, the 69 participants in our dataset stayed under the field of view of the cameras for different intervals of time. This implies that some acceleration streams from video will be partially or totally missing. Nevertheless, our method can also work in such cases, as can be seen in the formulation in Equation 2. The following is the empirical proof of this claim. Our experiments only consider missing video streams, but the insights found will also applied for missing streams from the wearable devices.

For each subset of participants in Table I (number of people under the field of view of the cameras for at least a given amount of time X), we applied the greedy and Hungarian association assignments without grouping. Those streams with less information than 10 minutes, for all subsets, were filled with zeros for the missing parts for practical purposes. Nonetheless, as explained in Section IV-B, our method will ignore these parts of the streams. Fig. 9 summarizes the association accuracy for these subsets. The first and last value of the plots are equal to those in Table III.

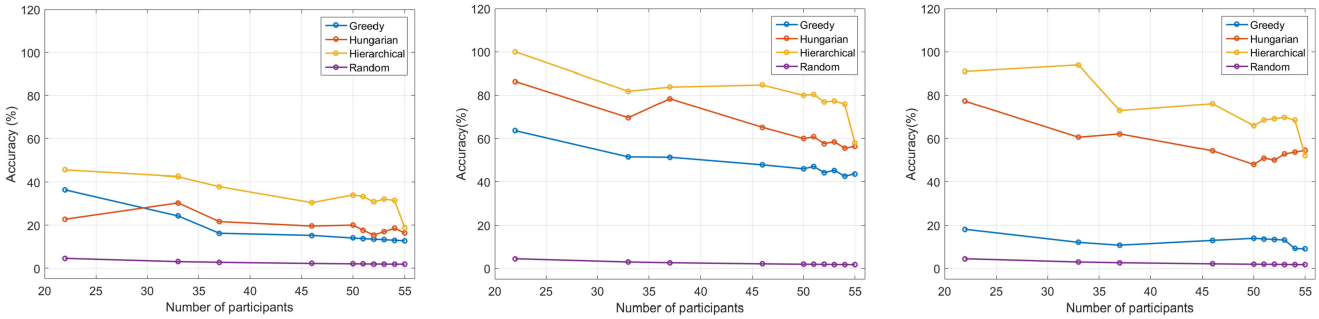

Fig. 10. Accuracy of stream association for our proposed method (Hierarchical), the state-of-the-art (Hungarian and Greedy) and the random baseline using (a) MI, (b) COV and (c) DTW as similarity metrics.

TABLE IV
ASSOCIATION ACCURACY AND NUMBER OF CORRECTLY ASSOCIATED GROUPS USING THE HIERARCHICAL HUNGARIAN METHOD FOR THE DIFFERENT SUBSETS OF PARTICIPANTS IN TABLE I(10 MINUTES INTERVAL)

					Accuracy (%)					
Num.	Num. groups**	TP_{group}	TP_{group}	TP_{group}	Hierarchical Hungarian			Ideal Hungarian(*)		
Partic.	$(min(p,q))$	(MI)	(COV)	(DTW)	MI	COV	DTW	MI	COV	DTW
22	16	9	16	14	45.45	100.00	90.91	68.18	100.00	90.91
33	25	11	19	23	42.42	81.82	93.94	78.79	100.00	93.94
37	28	10	22	20	37.84	83.78	72.97	81.08	100.00	89.19
46	34	10	27	25	30.43	84.78	76.09	69.57	100.00	91.30
46	34	10	27	25	30.43	84.78	76.09	69.57	100.00	91.30
50	37	12	27	23	34.00	80.00	66.00	72.00	100.00	92.00
51	38	12	28	25	33.33	80.39	68.63	72.55	100.00	92.16
52	38	13	29	25	30.77	76.92	69.23	73.08	100.00	92.31
53	38	14	29	26	32.08	77.36	69.81	73.58	100.00	92.45
54	39	14	30	26	31.48	75.93	68.52	74.07	100.00	92.59
69	40	10	31	26	18.84	57.97	52.17	57.97	88.41	81.16

*Results using the ground true groups and their correct matching (all manually annotated).
**A singleton is also treated as a group if output by the group detection as such.

Similar to the results on Table III, the combination of the COV as similarity metric and the Hungarian assignment has the best performance. Notice how the overall association accuracy decreases as the data becomes more incomplete and the number of participants increases.

Although there is an influence of the number of participants on the accuracy decrease, we believe this is strongly related to the missing data in the sets used as we gave a rather long observation interval. Nonetheless, for a subset of 51 people only present under the FoV of the cameras for 4 minutes (from the total 10 minutes), the normal hungarian method is still capable of matching correctly 60% of the streams. This can be further improved using our hierarchical method, as we see in the next subsection.

D. Evaluation of Group-to-Group Assignment

After analyzing the different components that can influence the association, we now introduce our Hierarchical Hungarian method which applies grouping.

Table IV summarizes the results for the association accuracies using this method. As well as in Fig. 9, for these associations we selected those participants that were under the field of view of the cameras for at least a given amount of time X, an calculated the association accuracy for these different subsets. This table also includes the total number of groups involved in the association. The last 3 columns of Table IV represents the accuracy with an ideal grouping. This means that the group formation of the participants (both in video and in the devices) are used and the correct group-to-group assignments $\{(e=1,f=1),(e=2,f=2),(e=3,f=3)\}$ are known. Thus, the the overall accuracy will be the mean accuracy for all $acc_{group}(e,f)$.

In addition, Fig. 10 show the association accuracy against the number of participants (as in Table I) for the 3 different metrics and the random baseline. Notice that both Table IV and Fig. 10 have sets with missing data.

Overall, all approaches are better than a random baseline (see Fig. 10). Furthermore, our Hierarchical method over-performs all other approaches when using the covariance as metric. Moreover, when analyzing the TP_{group} (groups correctly matched) of each association one can see that the association errors come from incorrectly matched groups in different modalities. For example, in the second row of Table IV we see that from 25 groups (in each modality, 50 in total) our method correctly matched 19, resulting in an accuracy of 81.82%. If all groups were correctly associated (ideal case), we can obtained a 100% accuracy using this metric, as seen in the second to last row. This implies that better algorithms to detect and match groups will improve our method. However, the correct group detection in each modality is not the main goal of this paper. Nonetheless, we proved that

TABLE V
COMPARISON OF OUR IMPROVED HIERARCHICAL METHOD TO OUR PREVIOUS VERSION PRESENTED IN [14] USING THE COVARIANCE (COV) AS METRIC

Method	Num. Participants							
	22	33	37	46	52	53	54	69
Acc. [6]	81.82	48.48	48.65	52.17	46.15	45.28	44.44	26.09
New	100	81.82	83.78	84.78	76.92	77.36	75.93	57.97

using group detection as a prior, even when defective, increases the association performance.

Comparison with previous work: To further evaluate the difference between our 2 implementations, we proceed to compare the results here presented to those obtained using the method presented in [14]. As explained in Sections IV-B and IV-C3, our method was optimized to account for missing data, either completely (missing streams) or partially (streams with missing data).

Table V summarizes the results for both methods using the covariance (COV) as metric, as this gave us the best results for both methods. Here, the sections with missing data in the streams were set to zero in order to use our old method. Nonetheless, our improved method account for this sections differently as was explained in Section IV-B.

The results for the complete set (22 people with no missing data) are rather similar with each other, and to what was presented in [14]. Here, as there is no missing data, the matrices for both methods are the same, which leads to the same result for the no grouping assignment (see Table III). The difference between the two is due to unmatched singletons, which remained after choosing an uneven group-to-group matching (each modality grouped the streams differently) and were omitted by our previous method. The improved method compensated for this issue.

In contrast, the results between both methods differ significantly as more missing data is introduced. These differences are due to 1) the way the values in the similarity matrices are calculated, and 2) the singletons omitted (and so unmatched) after an imperfect group-to-group association in our previous method. For example, for the case where only 60% of the streams are guaranteed (46 participants), 22 streams have complete data while the rest have different proportions of missing segments. While these segments are omitted when calculating the similarity matrices by our new method, they remained as zeros for [14] resulting in different values in the similarity matrices $\mathbf{D}$, and subsequently generating different values in the matrix of costs $\mathbf{C}$. Moreover, the latter can even result in a different group-to-group assignment. Nonetheless, as seen by these results, the improvements made to our hierarchical method account for such cases and maintain the functionality of our method for missing data.

E. Association vs. Social Context

The results obtained so far show that, although the length of the interval, the number of participants and amount of missing data have a significant impact on the accuracy, there are some

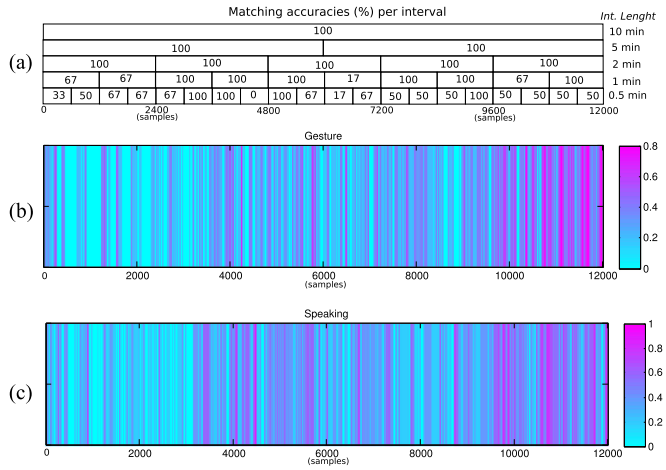

Fig. 11. Analysis of impact the impact of social actions in the association (better seen in color). (a) Matching accuracies for selected 6 participants under different lenght intervals at different times. (b) Normalized density of *hand gestures* for all participants (1 equals all participants gesturing). (c) Normalized density of *speaking* for all participants (1 equals all participants speaking).

confusions that cannot be totally explained by the aforementioned and detailedly described parameters.

We hypothesize that such confusions are due to the role of social context and in this section we intend to analyze this aspect further. To do so, we used social actions annotations provided with the MatchNMingle dataset [24] and specify on Section III. From all 8 social actions provided, we focus in hand gestures and speaking which are more related to conversational aspects of the context.

Fig. 11(a) shows the percentage of correct associations over time for 3 pairs of participants (6 people). These person stayed together for over an 90% of the 10 minute interval, so they have a high number of shared social actions. We selected 6 people as this is within the higher number of people interacting in the same group.

To obtain this figure, we took into consideration different interval lengths over time, so we can see the association performance for these 6 people for different times and observations lengths. Thus, the block in the far bottom right represents the accuracy percentage for the last interval of 0.5 minutes (600 samples) within our 10 minutes. Similarly, the top block represents the accuracy performance using the entire 10 minutes. Moreover, Figs. 11(b) and 11(c) represent the normalized density over time of the actions of hand gestures and speaking, respectively. With these figures, one can see graphically the correlation between social actions and the percentage of mismatches.

It can be seen from Fig. 11, specifically at the right side of all figures, that when there is a higher density of hand gestures and speaking (which are inherently associated with body movement [34]) the short intervals (bottom blocks of Fig. 11(a)) present a consistent lower association percentages compared to those where the occurrence of social actions is relatively low. This implies that shorter intervals with a high concentration shared of social actions become a failure case for our method. This

also relates with the trade-off between observation time and number of participants discussed in Section VI-B. Nonetheless, Fig. 11(a) also shows that even these cases can be compensated with longer observation intervals, to allow the method to properly discriminate between people interacting.

Furthermore, when analyzing the mismatches per individual we found that most mistakes are due to people talking to each other. So, for 2 people interacting actively (e.g. speaking and gesturing), our method switch their assignment, even within the same group. This also might explain why our hierarchical method is better than a normal Hungarian, as people moving at the same time but in different groups are not considered for an association.

VII. Discussion

As it was seen through Sections VI-A, VI-B, and VI-C; when applying our method without grouping, the performances of the association vary significantly with the metric and assignment used. Mainly, the Mutual Information (MI) performed poorly regardless of the assignment method, the Dynamic Time Warping distance (DTW) was competitive when using the Hungarian approach only, and using the covariance (COV) as metric gave us the best results for both assignment methods (greedy and Hungarian). This summary is better seen in Table III and Figs. 7, 8, and 9.

We hypothesize that the difference between the DTW and the COV lies on the local and global nature of the computation for each metric, respectively. The goal of the DTW is to warp one stream to the other optimally in time. Thus, the comparison between the streams is done locally up to some degree. In contrast, the COV takes into account the streams globally, computing implicitly the expected values of each entire stream separate and jointly.[1] From Fig. 8 we can see that the separation between the DTW and COV becomes smaller as the interval length for the observation reduces. For such cases, as the number of samples on each streams reduce, the two metrics start measuring similar distances. This also supports this global versus local hypothesis.

This analysis shows that not only the assignment with a global optimization is important. Also, a metric that computes the distances in a global manner is a better option for computing the distance matrices, specially for longer intervals of time. This might also explain why the DTW works only for the Hungarian method (a global optimization) but fails when using the greedy association (local).

A particularly interesting result is the low accuracy achieved when using the Mutual Information (MI) as similarity metric, as it is generally used for synchrony between streams. Even for a real scenario where the signals are more noisy (such as in our entire subset) the covariance and DTW distance are able to cope with the noise up to limit whereas the MI cannot. We hypothesis that this relates to what the MI is measuring in essence. This metric is more complex than just a measurement of similarity and, unlike the covariance and DTW distance, it is designed to also account for those moments of inverse synchrony. Hence, it might not be adequate for the association of the streams.

Another insight worth discussing is the difference between the methods' performances for different number of people to associate. We can see in Fig. 7 that the difference between the accuracy for 2 participants and 22 for complete streams (no missing data) is 12.9% for the Hungarian using the covariance, 19.7% for the Hungarian using DTW and 29.5% for the Greedy assignment with COV. These are our 3 best performing methods without using grouping yet.

We hypothesize that the difference comes with the global optimization performed by the Hungarian algorithm. Unlike other activities (e.g. walking, running), the actions performed during a conversation tend to be less discriminative and prone to missmatches even within a group, as seen in Section VI-E. This discrimination between streams becomes harder as the number of participants increases. Thus, a method that can optimize globally is preferred to handle this close nature of the interactions. This global optimization also applies for the metric used, as it was discussed at the beginning of this section. This might explain why the covariance by itself is performing acceptably with the greedy approach. Nonetheless, our divide-and-conquer approach analyzed in Section VI-D has proven to be a good alternative when the number of people increases.

VIII. Conclusion and Future Work

In this paper we showed a novel method for associating wearable devices to the person in the video wearing the device, using the acceleration and proximity data in both modalities. Our association method can handle missing data, either partially (incomplete streams) or completely (missing streams). This is fundamental as this method is intended for real crowded mingle scenarios where people can leave the scene at will (e.g. go to the bathroom) or the devices can fail.

Compared to previous efforts, we have significantly increased the number of devices used and regions in the video to associate. We also presented experiments to better understand the nature of this novel and relevant problem, focusing on the number of subjects to associate, length of the streams (time series) and proportion of missing data within the streams. These showed us that there is a strong relation between the number of people to associate and the length of the observations, in order to have enough information to make a discriminative assignment. Nonetheless, we have also found that using the spatial proximity as a prior for the associations significantly benefits the performance, even while associating rather short streams for several people. This is valid even when the group detection is imperfect in both modalities.

In our worst case scenario, where only a 31.9% of the streams in one of the modalities were complete and a 20% was entirely missing, our Hierarchical method manage to associate correctly 58% of the participants.

[1] $COV(X,Y)=E[XY]-E[X]E[Y]$, where X and Y are the 2 streams.

Our analysis of the mistakes made by our hierarchical Hungarian assignment showed us that these are due to the mismatch in the group-to-group assignment. Moreover, if the group-to-group were to be perfect (ground truth) the association accuracy will have increased significantly. Thus, future work should focus on methods for a better group detection in both modalities, and more efficient group-to-group matching. Perhaps a graph-to-graph matching will be a better option to the latter problem, adding structure to the group-to-group assignment.

Finally, we presented an analysis of the failure cases for our method, and how these are influenced by the social actions within a group. Thus, people sharing social actions (e.g. speaking or gesturing together) tend to be confused by the association method as their movement streams synchronize during their conversation.

APPENDIX A
LOCATION OF CAMERAS IN THE VENUE

Fig. 12 shows the location of the cameras in the venue, while collecting the MatchNMingle dataset. For the purposes of this work, we only focus on the mingle area.

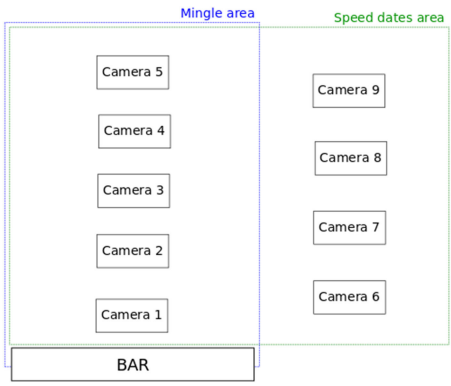

Fig. 12. Location of the cameras in the venue.

ACKNOWLEDGMENT

The authors would like to thank E. Gedik, L. van der Meij, and A. Demetriou for their collaboration and insights while collecting the data used in this work.

REFERENCES

[1] H. Hung and B. Krose, "Detecting F-formations as dominant sets," in *Proc. ACM Int. Conf. Multimodal Interact.*, 2011, pp. 231–238.
[2] S. Vascon *et al.*, "A game-theoretic probabilistic approach for detecting conversational groups," in *Proc. Asian Conf. Comput. Vision*, 2014, pp. 658–675.
[3] T. Gan, Y. Wong, D. Zhang, and M. Kankanhalli, "Temporal encoded F-formation system for social interaction detection," in *Proc. ACM Multimedia*, 2013, pp. 937–946.
[4] C. Martella, E. Gedik, L. Cabrera-Quiros, G. Englebienne, and H. Hung, "How was it? Exploiting smartphone sensing to measure implicit audience responses to live performances," in *Proc. ACM Multimedia*, 2015, pp. 201–210.
[5] X. Alameda-Pineda, Y. Yan, E. Ricci, O. Lanz, and N. Sebe, "Analizing free-standing conversational groups: A multimodal approach," in *Proc. ACM Multimedia*, 2015, pp. 5–15.
[6] W. Dong, B. Lepri, F. Pianesi, and A. Pentland, "Modeling functional roles dynamics in small group interactions," *IEEE Trans. Multimedia*, vol. 15, no. 1, pp. 83–95, Jan. 2013.
[7] D. Sanchez-Cortez, O. Aran, M. Schmid, and D. Gatica-Perez, "A nonverbal behavior approach to identify emergent leaders in small groups," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 816–832, Jun. 2012.
[8] D. Jayagopi, H. Hung, C. Yeo, and D. Gatica-Perez, "Modeling dominance in group conversations using nonverbal activity cues," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 3, pp. 501–513, Mar. 2009.
[9] A. Sapru and H. Bourlard, "Automatic recognition of emergent social roles in small group interactions," *IEEE Trans. Multimedia*, vol. 17, no. 5, pp. 746–760, May 2015.
[10] H. Hung and D. Gatice-Perez, "Estimating cohesion in small groups using audio-visual nonverbal behavior," *IEEE Trans. Multimedia*, vol. 12, no. 6, pp. 563–575, Oct. 2010.
[11] G. Bahle, P. Lukowicz, K. Kunze, and K. Kise, "I see you: How to improve wearable activity recognition by leveraging information from environmental cameras," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun.*, 2013, pp. 409–412.
[12] S. Stein and S. Mckenna, "Combining embedded accelerometers with computer vision for recognizing food preparation activities," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2013, pp. 729–738.
[13] 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.
[14] L. Cabrera-Quiros and H. Hung, "Who is where? Matching people in video to wearable acceleration during crowded mingling events," in *Proc. ACM Multimedia*, 2016, pp. 267–271.
[15] H. Hung and D. Gatica-Perez, "Estimating dominance in multi-party meetings using speaker diarization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 847–860, May 2011.
[16] F. Vallet, S. Essid, and J. Carrive, "A multimodal approach to speaker diarization on TV talk-shows," *IEEE Trans. Multimedia*, vol. 15, no. 3, pp. 509–520, Apr. 2013.
[17] 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.
[18] M. Rofouei, A. Wilson, A. Brush, and S. Tansley, "Your phone or mine? Fusing body, touch and device sensing for multi-user device-display interaction," in *Proc. ACM Conf. Human-Comput. Interact.*, 2012, pp. 1915–1918.
[19] A. Wilson and H. Benko, "CrossMotion: Fusing device and image motion for user identification, tracking and device association," in *Proc. Int. Conf. Multimodal Interact.*, 2014, pp. 216–223.
[20] T. Teixeira, D. Jung, and A. Savvides, "Tasking networked CCTV cameras and mobile phones to identify and localise multiple persons," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2010, pp. 213–222.
[21] O. Shigeta, S. Kagami, and K. Hashimoto, "Identifying a moving object with an accelerometer in a camera view," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2008, pp. 3872–3877.
[22] T. Plötz, C. Chen, N. Hammerla, and G. Abowd, "Automatic synchronization of wearable sensors and video-cameras for ground truth annotation - A practical approach," in *Proc. Int. Symp. Wearable Comput.*, 2012, pp. 100–103.
[23] L. Nguyen, Y. Kim, P. Tague, and J. Zhang, "IdentifyLink: User-device linking through visual and RF-signal," in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2014, pp. 529–539.
[24] L. Cabrera-Quiros, A. Demetriou, E. Gedik, M. v. d. L, 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. Affect. Comput.*, Jun. 25, 2018, to be published, doi: 10.1109/TAFFC.2018.2848914.
[25] C. Vondrick, D. Patterson, and D. Ramanan, "Efficiently scaling up crowdsourced video annotation," *Int. J. Comput. Vision*, 2012.
[26] L. Zhang and L. Van Der Maaten, "Structure preserving object tracking," in *Proc. IEEE Conput. Vision Pattern Recognit.*, 2013, pp. 184–204.
[27] T. Huynh and B. Schiele, "Analizing features for activity recognition," in *Joint Conf. Smart Objects Ambient Intell.: Innovative Context-Aware Services: Usages Technol.*, 2005, pp. 159–163.
[28] O. Lara and M. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Commun. Surveys Tut.*, vol. 15, no. 3, pp. 1192–1209, Jul.–Sep. 2013.

[29] E. Delaherche *et al.*, "Interpersonal synchrony: A survey of evaluation methods across disciplines," *IEEE Trans. Affect. Comput.*, vol. 3, no. 3, pp. 349–365, Jul.–Sep. 2012.
[30] B. Burkard, M. D.' Amico, and S. Martello, *Assignment Problems*. Philadelphia, PA, USA: SIAM, 2009.
[31] D. Pentico, "Assignment problems: A golden anniversary survey," *Eur. J. Oper. Res.*, vol. 176, no. 2, pp. 774–793, 2007.
[32] M. Cristani *et al.*, "Social interaction discovery by statistical analysis of F-formations," in *Proc. Brit. Mach. Vision Conf.*, 2011, pp. 23.1–23.12.
[33] C. Martella, M. Dobson, A. van Halteren, and M. Van Steen, "From proximity sensing to spatial-temporal social graphs," in *Proc. IEEE Int. Conf. Pervasive Comput. Commun.*, 2014, pp. 78–87.
[34] E. Gedik and H. Hung, "Personalised models for speech detection from body movements using transductive parameter transfer," *Pers. Ubiquitous Comput.*, vol. 21, no. 4, pp. 723–737, Aug. 2017.

Laura Cabrera-Quiros received the "Licenciatura" and the master's degree from the Instituto Tecnólogico de Costa Rica, Cartago, Costa Rica, in 2012 and 2014, respectively. She is currently working toward the Ph.D. degree in the Pattern Recognition and BioInformatics Group, Delft University of Technology, Delft, The Netherlands, working on automatic social behavior analysis using multimodal streams. In 2014, she received a full scholarship by the Costa Rican Government to pursue her postgraduate studies abroad. Her main interest is the use and fusion of wearable sensing and computer vision for applications oriented to analysis of social behavior.

Hayley Hung first degree in electrical and electronic engineering from Imperial College, London, U.K., and PhD in Computer Vision from Queen Mary University of London in 2007. She has been an Assistant Professor with the Pattern Recognition and Bioinformatics Group, TU Delft, Delft, The Netherlands, since 2013. Between 2010 and 2013, she held a Marie Curie Fellowship at the Intelligent Systems Lab, University of Amsterdam. Between 2007 and 2010, she was a Postdoctoral Researcher with the Idiap Research Institute, Switzerland. Her interests are social computing, social signal processing, computer vision, and machine learning.