Identification of Partitions in a Homogeneous Activity Group using Mobile Devices

Na Yu*, Yongjian Zhao†, Qi Han*, Weiping Zhu‡ and Hejun Wu†
* Department of Electrical Engineering and Computer Science, Colorado School of Mines, Golden, CO 80401
{nyu, qhan}@mines.edu
† Department of Computer Science, Sun Yat-sen University, Guangzhou, China
zhaoyj2@mail2.sysu.edu.cn, wuhejun@mail.sysu.edu.cn
‡ International School of Software, Wuhan University, Wuhan, China
cswpzhu@gmail.com

***Abstract*—People in public areas often appear in groups. People with homogeneous coarse-grained activities may be further divided into sub-groups depending on more fine-grained behavioral differences. Automatically identifying these sub-groups can benefit a variety of applications for group members. In this work, we focus on identifying such concurrent sub-groups. We present a generic framework using sensors built in commodity mobile devices. Specifically, we propose a two-stage process—sensing modality selection given a high level activity, followed by multimodal clustering to identify sub-groups. We develop one early fusion and one late fusion multimodal clustering algorithm. We evaluate our approaches using multiple datasets, two of them are with the same activity while the other has a different activity. The evaluation results show that the proposed multimodal-based approaches outperform existing work that uses only one single sensing modality and they also work in scenarios when manually selecting one sensing modality fails.**

I. Introduction

People often appear in groups and participate in various activities in public areas. People with homogeneous coarse-grained activities may further divide into sub-groups based on more fine-grained behavioral differences. For instance, tourists walk around in a park and walking is the same activity. Different walking flocks can be distinguished by the mobility patterns of the tourists, i.e., people in the same sub-group should have similar direction and speed. Another example is people watching a game. different subsets of the audience cheer for different teams in a game and the sub-groups can be distinguished by the specific actions performed by them, i.e., people in support of the same team typically perform certain gesture such as waving hands during the same time period when the team is performing well. Unobtrusively identifying these sub-groups can benefit various applications. For instance, a tour guide can better manage the tourist group, say, by sending customized "it is time to come back to the tour bus" alerts to different sub-groups depending on how far each group is from the bus; fans of the same team can be easily identified and information for future games and team information can be disseminated to the sub-groups. Partitioning groups with the same high level activities into sub-groups based on specific activity differences is exactly the focus of this work.

Lots of work has been done in individual or group activity recognition using mobile devices, but the problem at hand has not been fully addressed by existing work as detailed in Section II. We have been inspired by the divergence-based affiliation detection (DBAD) approach [1] which provides a framework to identify group affiliation given a sensing modality to be used for identifying an activity. Different from the group activity recognition problem which typically first recognizes each single user's activity and then analyzes their cooperative or collaborative relationship in a group [2], the group affiliation detection problem is about how to identify which users have similar behavior instead of identifying their specific activities. However, one limitation of DBAD is that only one sensing modality can be used at a time to distinguish multiple sub-groups, so it cannot accurately partition the groups when behavioral differences can be observed only through multiple sensing modalities. Another limitation of DBAD is that the sensing modality has to be explicitly provided to the framework, which is not practical in many cases since it is not clear what sensing modality works the best. In this work, we focus on building a generic framework that fuses multi-modal sensors to identify sub-groups in a homogeneous activity group based on fined-grained behavioral differences of the group members. In other words, the same high level activity of all the people is provided to the framework as prior knowledge, the framework will divide these people into sub-groups based on sensor readings of multiple sensing modalities.

Fine-grained partition of groups raises several interesting challenges.

- *Sensing modality selection.* Existing work has shown that sensors on the users' mobile devices produce similar signals when the users' activity is similar [3], therefore, group affiliation can be detected by monitoring the sensor signals of the mobile devices. However, with multiple sensing modalities available, it is not clear which subset of sensors can best capture users' activity similarity. It is even harder for a generic approach since it needs to detect group affiliation under any activity scenario. We address this issue in Section III.
- *Inconsistent window size among multiple sensing modalities.* To reduce cost (in particular in terms of energy

consumption) of data collection and exchange to measure similarity between users, it is necessary to summarize the sensor data time series into aggregate sensor features. we choose to use probability distribution function (PDF) as the aggregate sensor feature [1] . The length of sensor data time series for summarization significantly impacts similarity measurement, so we need to determine the measurement time window for each sensing modality and deal with the different time window sizes when combining the measurements of multiple sensing modalities. We address this issue in both training phase (Sections III-C) and testing phase (IV-A).

- *Multi-modal clustering.* Identifying groups based on the similarity measurements of multiple sensing modalities is non-trivial. Usually, we can apply clustering algorithms on the similarity graph of all users. However, since most sensing modalities are independent of each other, we cannot arbitrarily weigh each sensing modality to combine their similarity measurements into a single value. We address this issue in Section IV-B and Section IV-C.

Our approaches addresses these challenges in a generic framework using two phases: phase 1 is feature construction and phase 2 is multi-modal clustering. Further, we evaluate our approaches using both the dataset provided in DBAD and two datasets we collected. The evaluation results show that our proposed multimodal-based approach outperforms the DBAD approach that uses only one sensing modality by about 10% in group affiliation accuracy. Even though 10% is not a large margin, a distinguishing feature of our approach is that we can automatically select the right subset of sensing modalities while the best sensing modality has to be explicitly provided to DBAD, which significantly limits the practicality of DBAD. Further, our approach works effectively for various activity scenarios.

II. Related Work

Group affiliation detection and group identification have been studied using sensor-equipped mobile devices such as smartphones. There exist several ways to identify groups of people, for instance, based on interactions [4], proximity [5] [6], and activity [3] [1]. A homogeneous activity group refers to a group of people who perform similar activities and is a special case of activity-based groups. In addition, Grace [7] identifies groups of people who have face-to-face interactions, so it uses only wireless signal effects to identify the group relationships given that the group members are close to each other and have line of sight communications.

This work of identifying sub-groups in a homogeneous activity group is inspired by DBAD [1]. The DBAD approach uses a sensing modality represented as a histogram over a time window and delivers similar values for individuals with the similar activity. Specifically, it uses probability density functions (PDF) to model sensor data. Each mobile device computes the PDF of its local data and exchanges the PDF parameters with neighboring mobile devices. Then, each mobile device computes the disparity to its neighbors by computing the Jeffrey's divergence between the local PDF and the neighbors' PDFs. Further, the group affiliation between two mobile devices is determined by applying a threshold to the disparity value. The DBAD approach has several limitations. First, only one sensing modality is used at a time and this has to be selected manually. In particular, to identify people walking in different groups, the magnitude of the accelerometer readings is manually selected to identify groups walking with different speeds, and the azimuth sensing modality obtained from the orientation sensor is manually selected to identify groups with different walking directions. However, using only the azimuth will not work when different groups of people walk in the same direction but with different speeds; using only the magnitude can not differentiate groups with different directions. Therefore, multimodal sensing is necessary to distinguish different groups without prior knowledge of the grouping details. Second, in DBAD experiments, wearable mobile devices are attached to the human body with fixed positions to reduce noise in sensor data collected. This is not practical since people may put their phones in pockets or hold them in hand. It is not clear how DBAD performs when noise is present in the collected data.

In activity recognition, the first stage is often sensing modality selection (i.e., feature construction). There are many existing activity recognition approaches based on sensor-equipped mobile devices [8]. For examples, [9] presents a mobile application for real-time human activity recognition on the Android platform; [10] uses the accelerometer on mobile devices to recognize human activities including walking, jogging, ascending stairs, descending stairs, sitting, and standing; [11] uses the accelerometer on mobile devices to recognize hand gesture activities. In general, either based on some domain knowledge about the physical behavior involved or by making some default assumptions, a fixed set of sensing modalities is manually selected to construct the feature for a specific activity. Further, as discussed in [12], most activity recognition approaches are not generic and they often lead to solutions that are tied to the specific scenarios. Therefore, [12] proposes an algorithm which embeds feature construction into the machine learning process. Their algorithm can increase the feature search space, reduce the data preprocessing time, and is widely applicable. However, this generic approach only works for the classification and regression problems and cannot be directly applied to the clustering problem we face in this work.

III. Phase I: Sensing Modality Selection

For different activities, different sets of sensing modalities may be selected to represent the most distinguishing features of the activities. The sensing modality selection process uses a training set of sensor data for a given activity. The training set consists of one time series for each sensing modality. Each time series may have different sampling rate and may need to be summarized in different time windows. The goal is to select the sensing modalities which can provide accurate group affiliation detection results over the time windows. To achieve this goal, we first define scoring function to be used as a metric

to find the best window size for a sensing modality and then determine whether the sensing modality is qualified for group affiliation detection.

A. Scoring Function

We use a probability-based approach to predict the group affiliation detection accuracy of a sensing modality. Using the training set, with one PDF for each sensing modality on each mobile device, we can compute the Jeffrey's divergence [13] (which measures the disparity, opposite of similarity) between each device pair. The Jeffrey's divergence DJ between two probability distribution functions PDF_i and PDF_j is given by Eqn (1). It is an extension of the Kullback Leibler divergence, and it is numerically stable and symmetric [1].

$$DJ(PDF_i||PDF_j) = \int (PDF_i(m_k) - PDF_j(m_k)) ln(\frac{PDF_i(m_k)}{PDF_j(m_k)}) d(m_k) \tag{1}$$

Scoring function $F(m_k)$ (Eqn (2)) is defined as the probability of any pair of users in the n users' training set being in the same group when the Jeffrey's Divergence of them for sensing modality m_k is no larger than TH_s.

$$F(m_k) = P(G_{i,j} = 1 | DJ(PDF_i||PDF_j) \leq TH_s), \forall i, j \in n, i \neq j \tag{2}$$

where $G_{i,j} = 1$ indicates that i and j are affiliated with the same group while $G_{i,j} = -1$ indicates no group affiliation. TH_s is used to decide a pairwsie group affiliation in the test set. As discussed in [1], TH_s highly depends on the sensing modality being used and also varies for different activities. A practical value of TH_s can be experimentally obtained for each sensing modality in a specific activity scenario using datasets.

Using the Baye's theorem, Eqn (2) can be rewritten as Eqn (3):

$$\frac{P(DJ(PDF_i||PDF_j) \leq TH_s | G_{i,j} = 1) \times P(G_{i,j} = 1)}{\sum_{v=\{1,-1\}} P(DJ(PDF_i||PDF_j) \leq TH_s | G_{i,j} = v) \times P(G_{i,j} = v)} \tag{3}$$

The PDF of a sensing modality can be computed using Algorithm 1, where the distribution function type is known for the sensing modality. For example, most sensing modalities such as the 3D acceleration and the 3D rotation rate can be modeled as standard Gaussian distribution, and some sensing modalities such as the orientation data have circular features and can be modeled as von Mises distribution [14].

Algorithm 1 : Compute PDF

Input: time series s, time series length l, window size w, distribution function type f
Output: series of mixture model parameters p

1: **for** $i \in [0, l/w]$ **do**
2: Use expectation maximization [15] to calculate the parameters of f for values $s[i \times w]$ to $s[(i+1) \times w - 1]$;
3: $p[i] \leftarrow \{parameters\}$;
4: **end for**

B. Sensing Modality Selection

The sensing modality selection problem is stated as follows. Given n mobile devices or users in the training set, each has a set of time series S, one for each sensing modality under a given activity A, and given the scoring function F to predict the group affiliation detection accuracy, find the set of sensing modalities as well as the best window sizes which may result in an accuracy higher than decision threshold TH_d. Since the scoring function F defined in Eqn (3) is the probability that group affiliation is successfully detected, we compare F against a decision threshold TH_d to decide whether this probability may infer high accuracy in group affiliation detection. In general, TH_d should be larger than 0.5. This is because a probability less than 0.5 means that the group affiliation detection has more chance to be incorrectly detected than correctly detected. Further, according to different activity scenarios, TH_d may vary in order to choose the most significant sensing modalities which have highest scores. The determination of TH_d and the most significant sensing modalities will be discussed in Section V.

Algorithm 2 depicts how to select the candidate sensing modalities with their corresponding best window sizes which lead to the detection probability higher than TH_d.

Algorithm 2 : Select Sensing Modalities

Input: training set of time series $S_1, ..., S_n$ from n mobile devices under activity A, x sensing modalities in each set of time series, scoring function F, decision threshold TH_d
Output: set of candidate sensing modalities C

1: $C \leftarrow \emptyset$;
2: $score_{bestmodality} \leftarrow 0$;
3: $w_{bestmodality} \leftarrow 0$;
4: **for** $k \in [1, x]$ **do**
5: $m_k.index \leftarrow k$;
6: $m_k.score_{best} \leftarrow 0$;
7: $m_k.w_{best} \leftarrow w_{min}$;
8: **for** $w \in [w_{min}, w_{max}]$ **do**
9: **for** $i \in [0, n)$ **do**
10: $PDF[i] \leftarrow ComputePDF(s_i \leftarrow S_i[k], w)$;
11: **end for**
12: **if** $F(m_k) \geq score_{best}$ **then**
13: $m_k.score_{best} \leftarrow F(m_k)$;
14: $m_k.w_{best} \leftarrow w$;
15: **end if**
16: **end for**
17: **if** $score_{best} \geq TH_d$ **then**
18: $C \leftarrow C \cup \{m_k\}$;
19: **if** $score_{best} \geq score_{bestmodality}$ **then**
20: $score_{bestmodality} \leftarrow m_k.score_{best}$;
21: $w_{bestmodality} \leftarrow m_k.w_{best}$;
22: **end if**
23: **end if**
24: **end for**

C. Adjusting Window Size

The sensing modality selection process identifies the best and a few secondary sensing modalities. The window size of each candidate sensing modality is compared against that of the best sensing modality. For any candidate sensing modality,

if the new scoring function when using the window size of the best sensing modality is still no smaller than TH_d, the window size of this sensing modality will be modified to the same as that for the best sensing modality; otherwise, it keeps the original window size. The rationale behind this trick is to produce the multimodal fusion results mainly based on the best sensing modality and the results are expected to be improved by considering the secondary sensing modalities. The purpose of this window size matching is to reduce the processing of different window sizes during multimodal clustering in Phase II. Algorithm 3 depicts this process of adjusting window size.

Algorithm 3 : Adjust Window Size

Input: training set of time series $S_1, ..., S_n$ from n mobile devices under activity A, scoring function F, decision threshold TH_d, set of candidate sensing modalities C
Output: C with adjusted window sizes

```
for m_c ∈ C do
    if m_c.score_best < score_bestmodality then
        for i ∈ [0, n) do
            PDF_i ←
            ComputePDF(S_i[m_c.index], w_bestmodality);
        end for
        if F(m_c) ≥ TH_d then
            m_c.score_best ← F(m_c);
            m_c.w_best ← w_bestmodality;
        end if
    end if
end for
```

IV. Phase II: Group Identification using Multimodal Clustering

Once we have determined a set of candidate sensing modalities along with their window sizes, the next task is to identify sub-groups whose members have high similarity in these sensing modalities. The multimodal clustering problem has commonly been treated using fusion—early or late fusion [16]. Early fusion combines the sensing modalities in a specific representation before the clustering process, while late fusion first applies the clustering process to each sensing modality separately and then combines the results from each sensing modality. According to the comparison in [17], the advantage of early fusion is that it requires one learning phase only, while the disadvantage is the difficulty to combine multiple sensing modalities in a common representation. Although late fusion avoids this issue, it has other drawbacks such as the expensiveness in learning since every sensing modality requires a separate learning phase and potential loss of correlation in multi-dimensional space. We believe that early fusion may outperform late fusion in certain scenarios, but not in others. Therefore, we investigate and compare two clustering approaches—probability-based clustering for early fusion and majority voting-based clustering for late fusion.

Before we discuss the two clustering algorithms, we need to explain how to deal with different window sizes among different sensing modalities selected.

A. Dealing with Inconsistent Window Size

We use the window size of the best sensing modality for group identification, so the best sensing modality delivers one pairwise group affiliation result in each time window of group identification, and the secondary sensing modalities deliver multiple or no results in such a time window. Figure 1 shows an example with time series of three candidate sensing modalities provided by a mobile device, where s_1 is for the best sensing modality m_1, and the window size w_1 of m_1 is used as the group identification time window. The window size of each sensing modality is the same on all mobile devices. Therefore, by collecting the information of all sensing modalities on all mobile devices, the best sensing modality m_1 delivers one pairwise group affiliation result in each of the w_1 windows, the secondary sensing modality m_2 (corresponding to s_2) delivers one or no result, and the secondary sensing modality m_3 (corresponding to s_3) delivers one or multiple results.

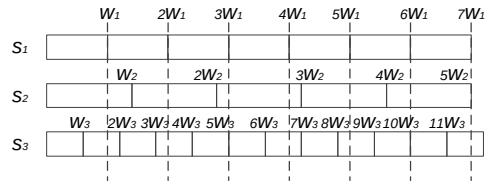

Fig. 1. Example time series with different window sizes

To determine pairwise group affiliation between a pair of mobile devices i and j, Jeffrey's divergence is compared against threshold TH_s: If $DJ(PDF_i||PDF_j) \leq TH_s$, then use the temporary result $v = 1$ to indicate positive group affiliation; otherwise, use $v = -1$ to indicate no group affiliation. Moreover, since the sensing modality m_k may deliver multiple results or no result in the group identification time window w_1, we define the aggregated result delivered by m_k in each group identification time window as $r_{m_k} \in \{1, 0, -1\}$, indicating whether the sum of v during the w_1 window is positive, zero, or negative. This is because positive summation implies most of the time positive group affiliation is suggested and vice versa.

Therefore, in each group identification time window, the sensing modality m_k delivers an aggregated result with value 1 indicating that a pairwise group affiliation is suggested, -1 indicating not suggested, and 0 indicating no result at this time. The aggregated result 0 may be caused by no result delivered in this time window or multiple results canceling out each other. In this case, the impact of sensing modality m_k on group identification does not need to be considered. Therefore, sensing modality m_k is taken into account in a group identification time window only when it provides an aggregated result 1 or -1. This consideration will be applied to the following clustering approaches.

B. Early Fusion: Probability-Based Clustering

In this section, we present an early fusion multimodal clustering approach which combines the pairwise group affiliation results delivered by all sensing modalities in each

group identification time window into a single result. A common approach for early fusion is to assign weights to each sensing modality. However, it is difficult to determine the appropriate weights, either manually or using a search procedure. Moreover, we have sensing modalities which deliver the pairwise group affiliation results with different accuracies. Intuitively, the best sensing modality should be given the highest weight in the early fusion process. If we assign a percentage as the weight to each sensing modality and then sum them up, the fusion function has no physical meaning and it is even more confusing than using only the best sensing modality. On the other hand, as discussed in Section II, using a single sensing modality without prior knowledge of grouping details is insufficient for many scenarios such as different groups of people walk in the same direction but with different speeds. Therefore, instead of using a single sensing modality or arbitrarily providing weights to different sensing modalities, we use the joint probability of correct pair-wise group affiliation detection as a fusion method to combine the pair-wise group affiliation results delivered by all the selected sensing modalities.

In a group identification time window, given a set of sensing modalities $\{m_1, ..., m_z\}$, each delivers a pairwise group affiliation result $r_{m_y} \in \{1, -1\}$, where $y \in \{1, ..., z\}$. The probability of correct pair-wise group affiliation detection (i.e., the fusion function) is calculated as shown in Eqn (5).

$$P(G_{i,j} = 1 | r_{m_1}, ..., r_{m_z}) = \frac{P(r_{m_1}, ..., r_{m_z} | G_{i,j} = 1) \times P(G_{i,j} = 1)}{\sum_{v=\{1,-1\}} P(r_{m_1}, ..., r_{m_z} | G_{i,j} = v) \times P(G_{i,j} = v)} \quad (5)$$

Further, we assume that each sensing modality can deliver a pairwise group affiliation result independently, so we can rewrite Eqn (5) as Eqn (6).

$$\frac{(\prod_{y=1}^{x} P(r_{m_y} | G_{i,j} = 1)) \times P(G_{i,j} = 1)}{\sum_{v=\{1,-1\}} (\prod_{y=1}^{z} P(r_{m_y} | G_{i,j} = v)) \times P(G_{i,j} = v)} \quad (6)$$

where the probabilities $P(r_{m_y} | G_{i,j} = v)$ and $P(G_{i,j} = v)$ are computed in the same way as the calculations in Section III-A using the training set. These pre-computed probability values can be directly applied to the clustering algorithm in which the test set is being used for group identification.

Using the test set, we can compute the pairwise group affiliation probabilities $P(G_{i,j} = 1 | r_{m_1}, ..., r_{m_z})$ in each group identification time window. We use a probability threshold TH_p to convert the pairwise group affiliation probabilities into a binary matrix V of the fused pairwise group affiliation results. The value corresponding to the mobile devices i and j in the matrix V is denoted as $V_{i,j} \in \{1, -1\}$. If $P(G_{i,j} = 1 | r_{m_1}, ..., r_{m_z}) \geq TH_p$, then $V_{i,j} = 1$, otherwise $V_{i,j} = -1$. Note that, we use -1 instead of 0 in order to be consistent with the value $v \in \{1, -1\}$ defined in Section IV-A. Similar to the discussion of the decision threshold TH_d in Section III-B, the probability threshold TH_p may vary for different activity scenarios. The determination of TH_p will be discussed in Section V.

Based on the group affiliation matrix, we can use existing clustering algorithms in one-dimensional space. We apply the density joint clustering algorithm (DJ-Cluster) [18] which is used by existing work of pedestrian flocks detection [6] to cluster the mobile devices into different groups. The overall process of the probability-based clustering approach is given in Algorithm 4. Note that, a sensing modality m_k is taken into account when computing the fused pairwise group affiliation result only when it provides the result $r_{m_k} \neq 0$.

Algorithm 4 : Probability-Based Clustering Algorithm

Input: test set of time series $S_1, ..., S_n$ on n mobile devices under activity A, x selected sensing modalities in each set of time series, probability threshold TH_p

Output: device groups in each group identification time window

```
Each mobile device uses its local time series to compute the PDFs for each selected sensing modality according to its window size;
The server or sink node collects the PDFs from all the n mobile devices once in each group identification time window and run the following process:
Initialize group affiliation matrix V;
for each device pair (i, j) do
    M ← ∅;
    for k ∈ {1, ..., x} do
        Compute r_{m_k};
        if r_{m_k} ≠ 0 then
            M ← M ∪ {(k, r_{m_k})};
        end if
    end for
    Compute p ← P(G_{i,j} | ∀r_{m_k} ∈ M);
    if p ≥ TH_p then
        V_{i,j} ← 1;
    else
        V_{i,j} ← −1;
    end if
end for
Apply DJ-Cluster algorithm on matrix V;
Return the clusters;
```

C. Late Fusion: Majority Voting-Based Clustering

In this section, we present a late fusion multimodal clustering approach which combines the clusters generated by each sensing modality in each group identification time window. We first use the DJ-Cluster algorithm to generate the clusters for each sensing modality separately. Similar to Algorithm 4, a sensing modality m_k is taken into account in the final cluster determination for two mobile devices only when it provides the result $r_{m_k} \neq 0$. We modify the majority voting approach used in [6]. The fusion by majority voting is basically calculating the sum of the weight of the sensing modalities where a pair of mobile devices are clustered into the same group. The two mobile devices are added as a cluster in the majority solution if the summed weight is larger than 50%. If one of the two mobile devices is already inside a solution cluster, the other one joins the same cluster instead of adding a new cluster. However, in [6], it simply assigns a weight of 50% to the features which may give the best accuracy and then divide the remaining 50% among the other features. It does not search for the best weights assignment or automatic training

of these weights. Therefore, the weight assignment is still a problem in this late fusion multimodal clustering approach. Since we already have a sensing modality selection process before the clustering process, as long as the sensing modalities are well selected, all the selected sensing modalities should play important roles in the group identification. Therefore, we apply the same weight on all selected sensing modalities. Algorithm 5 gives the overall process of the majority voting-based clustering approach.

Algorithm 5 : Majority Voting-Based Clustering Algorithm

Input: test set of time series $S_1, ..., S_n$ on n mobile devices under activity A, x selected sensing modalities in each set of time series
Output: device groups in each group identification time window

1: The same as line 1 in Algorithm 3;
2: The same as line 2 in Algorithm 3:
3: **for** each device pair (i, j) **do**
4: $M_{i,j} \leftarrow \emptyset$;
5: **end for**
6: **for** $k \in \{1, ..., x\}$ **do**
7: Initialize group affiliation matrix V;
8: **for** each device pair (i, j) **do**
9: Compute r_{m_k};
10: **if** $r_{m_k} \neq 0$ **then**
11: $M_{i,j} \leftarrow M_{i,j} \cup \{k\}$;
12: $V_{i,j} \leftarrow r_{m_k}$;
13: **else**
14: $V_{i,j} \leftarrow -1$;
15: **end if**
16: **end for**
17: Apply DJ-Cluster algorithm on matrix V;
18: **end for**
19: **for** each device pair (i, j) **do**
20: Apply majority voting to the clusters generated by the sensing modalities in $M_{i,j}$;
21: **end for**
22: Return the final clusters in the majority solution;

V. Performance Evaluation

In performance evaluation, we first use the dataset provided in DBAD [1] where the activity is people walking together. There are two limitations of the DBAD dataset as discussed in Section II: one is that the wearable mobile devices are attached to the human body with fixed positions in order to reduce noise in the collected sensor data; the other is that there is only one activity (i.e., people walking together) involved. Therefore, we also collect our own datasets—one for the park scenario and one for the game scenario as discussed in Section I. The park scenario has the same activity with the DBAD dataset, but with less controlled phone positions to allow for more noisy data and with more sensor modalities to allow for consideration of multiple modalities. The game scenario has a different activity (i.e., audience wave hands for different teams) from the DBAD dataset and it is used to demonstrate that our approaches are general and can handle different activities. Then, we test our approaches on these datasets. For each dataset, we divide it into two parts—the first half as the training set for sensing modality selection and the second half as the test set for identification of sub-groups within a homogeneous activity group. We implement our algorithms in Python and run Algorithms 2 and 3 on the training set while Algorithms 4 and 5 on the test set.

Since the DBAD approach only detects pairwise group affiliation, its evaluation only considers the accuracy of pairwise group affiliation detection results. In contrast, the final results of our approaches are the identified groups, therefore we use the performance metrics pairwise group affiliation accuracy and group membership similarity to evaluate the intermediate and the final results, respectively. For group identification, since the groups are pre-configured and unchanged during an experiment, we determine the final groups when the grouping results are stable, i.e., groups remain for at least five group identification time windows. The group membership similarity is calculated as the average Jaccard similarity [19] between an identified group and the corresponding actual group. The pairwise group affiliation accuracy is calculated as radio of the number of correctly determined group relationships over the total number of pairwise group relationships when the final groups are identified.

A. Results using the DBAD Dataset

The DBAD date set contains the sensor data obtained from 10 homogeneous Android devices which are attchahed on the hip of each person. The experiments are conducted with different group configurations (from 1 group to 10 groups), and each experiment lasts 51 minutes. Data is captured from 3D accelerometer and orientation sensor. To compute the activity similarity for people walking together, we consider the following sensing modalities available in the dataset: x-acceleration, y-acceleration, z-acceleration, magnitude (obtained from the 3D accelerometer); azimuth, pitch, roll (obtained from the orientation sensor). The magnitude measurement is the square root of the square sum of the 3D acceleration measurements, and the DBAD evaluation uses it instead of the 3D acceleration measurements.

We apply Algorithms 2 and 3 on the training set. We set the minimum and maximum window sizes as 5 seconds and 50 seconds, respectively. Table I shows the results for each sensing modality, where the best score is the scoring function with the best window size for that sensing modality and the new score is the recalculated scoring function using the best sensing modality's best window size.

TABLE I
Sensing Modality Selection using DBAD Dataset

Sensing modality	Best window size	Best score	New score
x-acceleration	15 s	0.65	0.5
y-acceleration	15 s	0.64	0.5
z-acceleration	15 s	0.55	0.49
magnitude	15 s	0.58	0.5
azimuth	**5 s**	**0.75**	**0.75**
pitch	5 s	0.45	0.45
roll	5 s	0.48	0.48

As discussed in Section III-B, the decision threshold TH_d should be larger than 0.5. Here we set the decision threshold

$TH_d = 0.55$, then the azimuth (window size 5 s) , x-acceleration (window size 15 s), y-acceleration (window size 15 s), z-acceleration (window size 15 s), and magnitude (window size 15 s) are selected. Since magnitude is a redundant sensing modality to the 3D acceleration and it yields very similar score as the 3D acceleration, we use the 3D acceleration sensing modalities in Algorithms 4 and 5 instead of magnitude. We next use the test set to evaluate Algorithms 4 and 5.

First, we consider the probability threshold TH_p in Algorithm 4. Similar to the decision threshold TH_d, it should also be larger than 0.5. Therefore, we vary it from 0.55 to 0.95. Fig. 2 shows the results of both the pairwise group affiliation accuracy and the group membership similarity. The group membership similarity is slightly smaller than the pairwise group affiliation accuracy. This is because that there exist some critical links in the graph-based clustering algorithms. If a critical link is determined with incorrect group affiliation result, it will significantly impact the group identification results. In general, the pairwise group affiliation accuracy increases when TH_p increases. Using the DBAD dataset, $TH_p = 0.85$ leads to both the highest pairwise group affiliation accuracy and the highest group membership similarity. Next, we will compare the results of the probability-based clustering algorithm using $TH_p = 0.85$ with the results of using the DJ-Cluster algorithm on each single sensing modality as well as using the majority voting-based clustering algorithm on all sensing modalities.

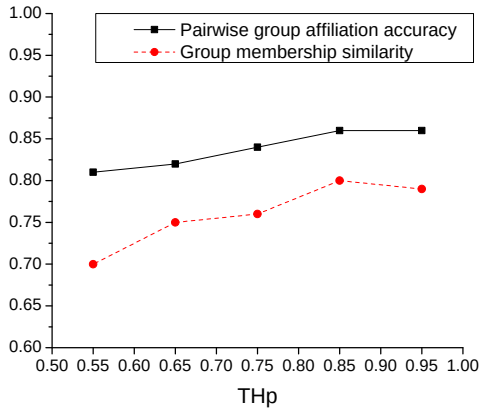

Fig. 2. Impact of TH_p using DBAD dataset

Fig. 3 shows the pairwise group affiliation accuracy using different sensing modalities or different approaches. Since the majority voting-based clustering algorithm outputs the final clusters based on the clusters computed from each sensing modality, it does not output the combined pairwise group affiliation results of all sensing modalities, we only compare the probability-based approach with each single sensing modality for the pairwise group affiliation accuracy. The 3D acceleration sensing modalities lead to an accuracy around 0.6 while the azimuth related to the orientation sensor leads to an accuracy about 0.76. These results are consistent with the findings in the DBAD approach, where the azimuth delivers the best pairwise group affiliation accuracy. Beyond their findings, our sensing modality selection approach automatically selects the azimuth as the most significant sensing modality. Further, the probability-based approach leads to an accuracy about 0.86, which shows that the multimodal-based approach proposed in this paper outperforms the original DBAD approach which uses a single sensing modality.

Fig. 4 shows the group membership similarity, where the comparisons are similar to Fig. 3. In addition, the majority voting-based approach provides a higher group membership similarity than using the 3D acceleration or the azimuth separately. This again shows that the multimodal-based approach is better than using a single sensing modality. Moreover, the probability-based approach outperforms the majority voting-based approach using the DBAD dataset.

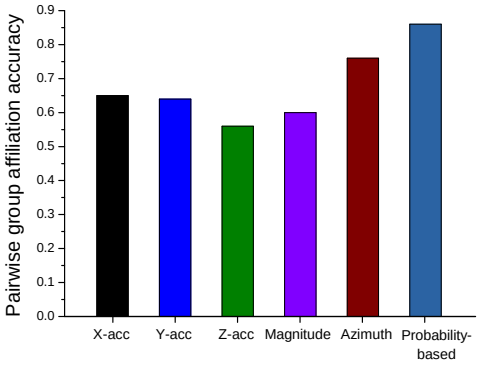

Fig. 3. Pairwise group affiliation accuracy using DBAD dataset

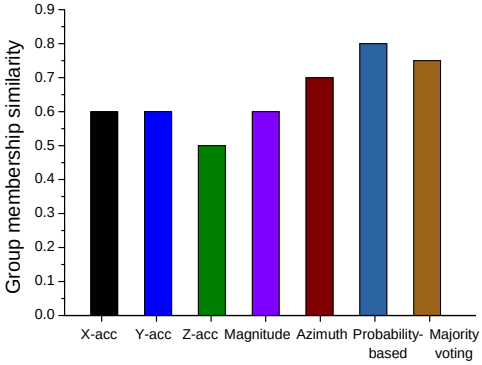

Fig. 4. Group membership similarity using DBAD dataset

B. Results using the Park Scenario Dataset

Since the DBAD dataset only contains accelerometer and orientation sensor, we collect our own dataset with more motion sensors on smartphones for the same activity that people walk together. It contains the sensor data obtained from 8 heterogeneous smartphones (e.g., Nexus and Samaung Galaxy phones) held in hands by people walking in 3 groups for about 10 minutes. These groups have different walking directions and are slightly different in walking speed. The sensors recorded are 3D accelerometer, 3D gyroscope, and orientation sensor. We consider the following sensing modalities: x-acceleration, y-acceleration, z-acceleration (obtained from the 3D accelerometer); x-rotation, y-rotation, z-rotation (obtained

from the 3D gyroscope); azimuth, pitch, roll (obtained from the orientation sensor).

We apply Algorithms 2 and 3 on the training set and use the same minimum/maximum window sizes as in the DBAD training set. Table II shows the results, where the azimuth also leads to the best score as in Table I.

TABLE II
SENSING MODALITY SELECTION USING PARK SCENARIO DATASET

Sensing modality	Best window size	Best score	New score
x-acceleration	15 s	0.58	0.45
y-acceleration	15 s	0.55	0.45
z-acceleration	15 s	0.51	0.4
x-rotation	15 s	0.42	0.4
y-rotation	15 s	0.35	0.33
z-rotation	15 s	0.35	0.33
azimuth	**5 s**	**0.78**	**0.78**
pitch	5 s	0.4	0.4
roll	5 s	0.4	0.4

We also choose the decision threshold $TH_d = 0.55$, so the azimuth (window size 5 s), x-acceleration (window size 15 s), and y-acceleration (window size 15 s) are the selected sensing modalities. Although z-acceleration is not selected here, it does not contribute significant results in Section V-A either. Fig. 5 shows the the results of the probability-based approach when we vary the probability threshold TH_p from 0.55 to 0.95. Similar to the findings in the DBAD test set, the group membership similarity is slightly lower than the pairwise group affiliation accuracy, and the pairwise group affiliation accuracy increases when TH_p increases. We choose $TH_p = 0.85$ for the probability-based approach in the following comparisons using the test set.

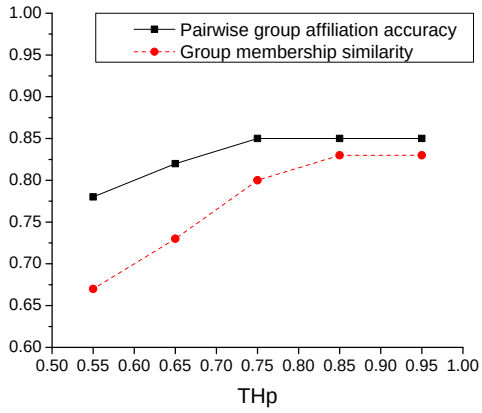

Fig. 5. Impact of TH_p using park scenario dataset

Fig. 6 compares the pairwise group affiliation accuracy results. Similar to the results in Fig. 3, the azimuth leads to a higher accuracy than the 3D acceleration, and the probability-based approach leads to an even higher accuracy than the azimuth. Fig. 7 compares the group membership similarity results. The comparison is consistent with that of the pairwise group affiliation accuracy. In addition, the majority voting-based approach leads to a lower group membership similarity than the probability-based approach, but the similarity is still higher than using the x-acceleration, y-acceleration, or azimuth individually. All these results again verify that the multimodal-based approaches outperform the original DBAD approach that works with a single sensing modality. Further, unlike the controlled experiments with homogeneous phones and fixed phone positions in the DBAD approach, our experiments are less controlled and have more uncertainty in the collected sensor data. Despite all these, the results using our dataset are still promising (e.g., the group membership similarity for the probability-based approach is still above 0.8), indicating that our approaches can inherently deal with sensor data noises. This is because sensing modalities are selected in the presence of data noises.

Moreover, the results using the park scenario dataset are consistent with those using the DBAD dataset because of the same activity involved. This indicates that the same training set for the same activity may be used to test both the datasets if the training set is well collected and the parameters involved in the algorithms are well studied.

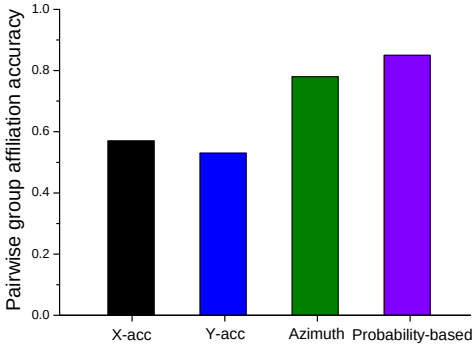

Fig. 6. Pairwise group affiliation accuracy using park scenario dataset

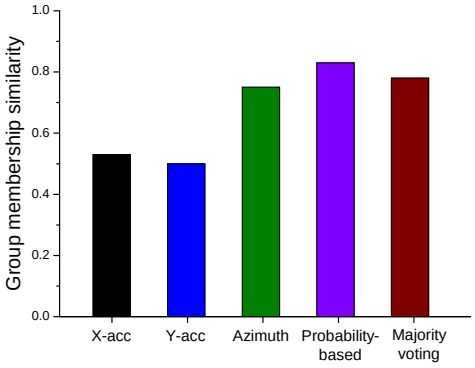

Fig. 7. Group membership similarity using park scenario dataset

C. Results using the Game Scenario Dataset

We also collect our own dataset for the game scenario discussed in Section I. This dataset also contains the sensor data obtained from 8 heterogeneous smartphones for about 10 minutes. Each group waves their smartphones in different time periods, mimicking the activity that audience cheer for the two competitor teams in a game. The sensors recorded are the same as in the park scenario dataset. Similarly, we use the same settings and apply Algorithms 2 and 3 on the

training set. Table III shows the results of sensing modality selection. Different from Table I and II, the 3D rotation lead to the highest scores. The 3D acceleration may still work, but the azimuth does not make much sense in this activity scenario. This implies that the DBAD approach of manually selecting one single sensing modality will not work in such a scenario.

TABLE III
SENSING MODALITY SELECTION USING GAME SCENARIO DATASET

Sensing modality	Best window size	Best score	New score
x-acceleration	15 s	0.66	0.66
y-acceleration	15 s	0.65	0.65
z-acceleration	15 s	0.58	0.58
x-rotation	15 s	0.75	0.75
y-rotation	**15 s**	**0.8**	**0.8**
z-rotation	15 s	0.72	0.72
azimuth	5 s	0.54	0.51
pitch	5 s	0.52	0.5
roll	5 s	0.46	0.45

We can still choose the decision threshold $TH_d = 0.55$, so the x-acceleration, y-acceleration, z-acceleration, x-rotation, y-rotation, and z-rotation are selected. Fig. 8 shows the results of the probability-based approach. Similar to the findings in both the DBAD test set and the park scenario test set, we can choose $TH_p = 0.95$ for the probability-based approach to compare with using each single sensing modality as well as the majority voting-based approach.

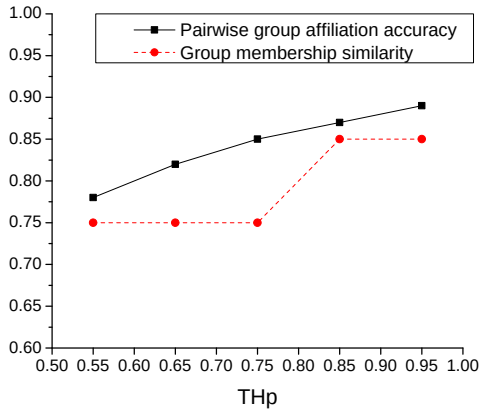

Fig. 8. Impact of TH_p using game scenario dataset

Fig. 9 shows that the y-rotation leads to a higher accuracy than each other sensing modality, and the probability-based approach leads to even higher accuracy than using only y-rotation. Fig. 10 shows a consistent trend as in Fig. 9. However, different from both Fig. 4 and Fig. 7, the majority voting-based approach leads to a slightly higher group membership similarity than the probability-based approach. This is because there are several significant sensing modalities (i.e., x-rotation, y-rotation, and z-rotation) which contribute accurate results in this activity scenario. Unlike the activity that people walk together, only the azimuth makes significant contribution in the final results of the multimodal-based approaches, here all the 3D rotation make significant contributions, therefore the majority voting is more significant.

In summary, the activity scenario significantly impacts the sensing modality selection as well as the group identification results. This verifies our hypothesis in Section III that a selection process is needed to automatically select sensing modalities for different activities. In addition, the comparison of the probability-based approach and the majority voting-based approach verifies our hypothesis in Section IV that the early fusion multimodal clustering approaches may outperform late fusion in some activity scenarios, but not always. All things considered, all the approaches proposed in this work (i.e., Algorithms 2, 3, 4, and 5) are effective for various activity scenarios.

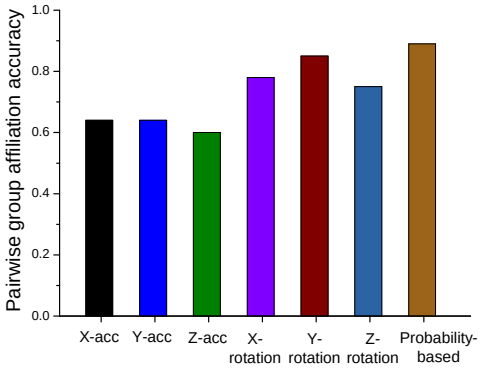

Fig. 9. Pairwise group affiliation accuracy using game scenario dataset

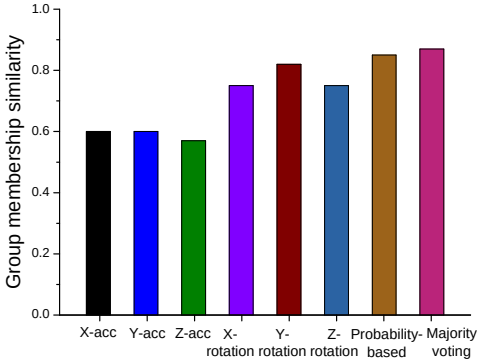

Fig. 10. Group membership similarity using game scenario dataset

VI. CONCLUSION

In this paper, we present a generic framework to identify sub-groups in a homogeneous activity group using sensor-equipped mobile devices. We have first proposed a sensing modality selection approach given a high level activity. We have then provided an approach to deal with the multiple window sizes among all the selected sensing modalities. By setting the group identification window size the same as that of the best sensing modality, we have further developed two multimodal clustering approaches—probability-based approach for early fusion and majority voting-based approach for late fusion. At last, we have evaluated our approaches using a publicly available dataset and also two others collected by ourselves. The evaluation results have shown that our framework of multimodal approaches outperforms the original

DBAD approach which works on a single sensing modality, and the framework is effective for various activity scenarios.

Several improvements are considered for future work. First, in this framework, activity is considered as an input to the algorithms. Although we have not yet studied the sensing modality selection training per activity, our evaluation results of different datasets but with the same activity tend to be very similar, indicating that using the same training set to build group identification models for an activity and test on different datasets regarding this activity is possible. Second, in this work, we assume that the sensor data distributions of all mobile devices are periodically (i.e., according to the window sizes of the sensing modalities) sent to a central server in an infrastructure-based environment or collected by a sink node via data collection protocols in mobile ad hoc networks. Therefore, the central server or the sink node has the complete information in the network to calculate pairwise similarities and apply clustering algorithms on the group affiliation matrix based on the pairwise similarities. In our future work, we will further consider a pure peer-to-peer environment where neighboring mobile devices exchange their sensor data distributions. Since some pairwise similarities between multi-hop neighbors may not be computed due to limited hops of data exchange, the clustering algorithms need to be revised accordingly to work with a local partial group affiliation matrix on each mobile device. Last, we will apply the Jeffrey's divergence directly to multiple sensing modalities when a practical mathematical method is available.

ACKNOWLEDGEMENT

This project is supported in part by NSF grant CNS-0915574.

REFERENCES

[1] D. Gordon, M. Beigl, M. Wirz, G. Troster, and D. Roggen, "Peer-to-peer group affiliation detection using mobile phones," in *ISWC*, 2014.

[2] D. Gordon, J. Hanne, M. Berchtold, A. A. N. Shirehjini, and M. Beigl, "Towards collaborative group activity recognition using mobile devices," *Mobile Networks and Applications*, pp. 326–340, 2013.

[3] D. Roggen, D. Helbing, G. Troster, and M. Wirz, "Recognition of crowd behavior from mobile sensors with pattern analysis and graph clustering methods," *Networks and Heterogeneous Media*, pp. 521–544, 2011.

[4] B. Guo, H. He, Z. Yu, D. Zhang, and X. Zhou, "Groupme: Supporting group formation with mobile sensing and social graph mining," in *Mobiquitous*, 2012.

[5] M. B. Kjargaard, M. Wirz, D. Roggen, and G. Troster, "Mobile sensing of pedestrian flocks in indoor environments using wifi signals," in *PerCom*, 2012.

[6] ——, "Detecting pedestrian flocks by fusion of multi-modal sensors in mobile phones," in *UbiComp*, 2012.

[7] N. Yu and Q. Han, "Grace: Recognition of proximity-based intentional groups using collaborative mobile devices," in *MASS*, 2014.

[8] O. D. Lara and M. A. Labrador, "A survey on human activity recognition using wearable sensors," *IEEE Communications Surveys and Tutorials*, 2013.

[9] ——, "A mobile platform for real time human activity recognition," in *CCNC*, 2012.

[10] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," in *SensorKDD*, 2010.

[11] J. Liu, Z. Wang, L. Zhong, J. Wickramasuriya, and V. Vasudevan, "uwave: Accelerometer-based personalized gesture recognition and its applications," in *PerCom*, 2009.

[12] R. Cachucho, M. Meeng, U. Vespier, S. Nijssen, and A. Knobbe, "Mining multivariate time series with mixed sampling rates," in *UbiComp*, 2014.

[13] Wikipedia, "Divergence (statistics)," http://en.wikipedia.org/wiki/Divergence_(statistics).

[14] S. Calderara, A. Prati, R. Reignier, and J. L. Crowley, "Mixtures of von miss distribution for people trajectory shape analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 457–471, 2011.

[15] Wikipedia, "Expectationmaximization algorithm," http://en.wikipedia.org/wiki/Expectationmaximization_algorithm.

[16] G. Petkos, S. Papadopoulos, E. Schinas, and Y. Kompatsiaris, "Graph-based multimodal clustering for social event detection in large collections of images," *MultiMedia Modeling*, pp. 146–158, 2014.

[17] C. Snoke, M. Worring, and A. Smeulders, "Early versus late fusion in semantic video analysis," in *MULTIMEDIA*, 2005.

[18] C. Zhou, D. Frankowski, P. Ludford, S. Shekhar, and L. Terveen, "Discovering personal gazetteers: an interactive clustering approach," in *GIS*, 2004.

[19] Wikipedia, "Jaccard index," http://en.wikipedia.org/wiki/Jaccard_index.