6-2016

Fusing WiFi and video sensing for accurate group detection in indoor spaces

Kasthuri JAYARAJAH
Singapore Management University, kasthurij.2014@phdis.smu.edu.sg

Zaman LANTRA
University of Moratuwa

Archan MISRA
Singapore Management University, archanm@smu.edu.sg

Citation

JAYARAJAH, Kasthuri; LANTRA, Zaman; and MISRA, Archan. Fusing WiFi and video sensing for accurate group detection in indoor spaces. (2016). *WPA '16: Proceedings of the 2016 Workshop on Physical Analytics, Singapore, 2016 June 26*. 49-54. Research Collection School Of Information Systems.
Available at: http://ink.library.smu.edu.sg/sis_research/3636

Fusing WiFi and Video Sensing for Accurate Group Detection in Indoor Spaces

Kasthuri Jayarajah, Zaman Lantra[†][*], Archan Misra
Singapore Management University
[†]University of Moratuwa
{kasthurij.2014, archanm}@smu.edu.sg, [†]zaman.lantra@gmail.com

ABSTRACT

Understanding one's group context in indoor spaces is useful for many reasons – e.g., at a shopping mall, knowing a customer's group context can help in offering context-specific incentives, or estimating taxi demand for customers exiting the mall. Group detection and monitoring using WiFi-based indoor location traces fails when users are invisible (either because they don't carry smartphones, or because their WiFi is turned OFF) or when location tracking is inaccurate. In this paper, we propose a multi-modal group detection system that fuses two independent modes: video and WiFi, for detecting groups with low latency and high accuracy. We present preliminary results from a micro-study with 20 group episodes and report an overall precision of 0.81 and recall of 0.9, an improvement of over ≈20% over WiFi-based group detection.

Keywords

Group Monitoring; Multi-Modal Sensing; Sensor Fusion

1. INTRODUCTION

Detecting the *group-context* (i.e., whether a person is alone or moving with additional individuals) of visitors to indoor venues (e.g., shopping malls, convention centers and airports) is useful for a variety of applications, including targeted recommendations & advertising, inference of ties between people [6] and visitor analytics [5]. Such group detection typically utilizes one of two approaches: *video analytics* [10] or *WiFi based movement analytics* [15]. Each approach is known to have its own limitations:

- Video-based group detection is not pervasive–typically, cameras monitor only parts of the indoor venue. Video-based approaches can also exhibit false positives (when unrelated individuals incidentally happen to traverse the observation region together). Most importantly, video analytics cannot *identify* specific individuals in a group (facial recognition is not applicable to random individuals visiting a public space), and is thus not amenable to applications such as targeted recommendation.
- While usually pervasive throughout the building, group detection based on WiFi based location data suffers from false negatives (individuals without a WiFi enabled device are effectively invisible to the group detector). Moreover, practical server-side WiFi localization is known to have errors around $\pm\ 6-8m$ and update latencies of several minutes, and can thus generate false conclusions in densely-crowded indoor spaces.

To tackle these limitations in a practical way, we ask the question: *Can we improve the accuracy of group detection by smartly fusing the sensing modalities of video (with only intermittent coverage) and WiFi (that provides pervasive location tracking)?* The fundamental idea is as follows: video-based analytics is used to identify the *number* of individuals that transit a given area "together" at a specific time instant. WiFi based group detection also uses such detected transitions, as well as periods of collocation at 'stay points' before/after the transition, to separately identify potential groups. By matching the video-based transition pattern with the WiFi based location-cum-residency pattern of potential groups, we hope to reduce both false positives and negatives, as well as tackle the problem of *hidden nodes* (individuals with no WiFi enabled devices). Note: this group detection problem is distinct from WiFi+video based *localization*: our goal is not to explicitly track the location, but instead to identify individuals who move together.

Challenges: A practical fusion approach must address three failure modes associated with isolated use of WiFi or video data: (a) *Hidden nodes:* As many users, such as the elderly and children may not carry a mobile device, they have no WiFi footprint. However, for improved accuracy, it is important to uncover the presence of such users, and associate them with partially-identified groups. (b) *Sub-group Identification:* Commercial server-based WiFi location systems (which provide the universal observability of all devices in the public venue without requiring any specific cooperation from users) often have high update latency (the gap between successive WiFi readings are often 3-4 mins). Consequently, such WiFi based systems are unable to separate out two groups of users who transit the same region in rapid succession (e.g., separated by a few seconds), and instead aggregate them into a common, large group. (c) *Location Errors:* As server-side WiFi readings are often delayed, two individuals entering a location concurrently may find location estimates be separated out by several minutes. Such temporal discrepancy may cause the group detector to fail.

In this paper, we provide early evidence that these challenges can be tackled by intelligently fusing the video and WiFi based group membership inferences. Using 20 curated episodes of group and

[*]Work done during an internship at Singapore Management University

WPA'16, June 25-30, 2016, Singapore, Singapore
© 2016 ACM. ISBN 978-1-4503-4328-2/16/06…$15.00
DOI: http://dx.doi.org/10.1145/2935651.2935659

individual movement behavior, we identify the distinct spatiotemporal variations that occur naturally in groups vs. individuals, and make the following **Key Contributions:**

- *Benchmark Video-based Group Detection:* We first develop and evaluate a state-of-the-art technique for video-based group detection, consisting of 3 steps: (a) localizing people within video frames, (b) tracking detected persons across frames and finally, (c) detecting groups at low-latency. We show that, for our movement episodes, the video-based detection reaches up to 80% precision and recall, at a latency of 0.5 seconds.

- *Sensor Fusion and Probabilistic Group Membership:* We propose and implement a candidate group detection algorithm that fuses such video and WiFi observations. The algorithm first computes the probability of group membership in each mode independently, using the duration of overlapping collocation periods for WiFi (i.e., using the *GruMon* technique [15]) and the "duration elapsed since last collocation" for video. It then combines these probabilities to provide more robust group detection, even in the presence of hidden nodes and sub-groups. With 33% of the nodes *hidden*, and 40% of the groups being *subgroups*, our candidate algorithm achieves a recall of 0.9 – this is a $\approx 20\%$ improvement over a WiFi-only approach, and a 10% improvement over the recall of a video-only approach. Likewise, the algorithm achieves a 30% improvement in precision over the WiFi-only approach. In particular, our algorithm detects 67% of hidden nodes and *all* subgroups, compared to 0% for WiFi.

2. MOTIVATING SCENARIOS

In Sen et. al. [15], we provide motivating application scenarios for group detection in multi-functional indoor spaces such as shopping complexes. Here, we provide additional scenarios where accurate detection is warranted.

Visitor Analytics for Surveillance: Public spaces such as train stations, airports, amusement parks, etc. experience high levels of flux of visitors. Longitudinal profiling of visitors, both individuals and groups, allows for detecting anomalies (e.g., an unusual congregation of people at a secure area). In recent works [5], it was shown that groups indeed behave differently than individuals– typically spending longer stay times with fewer transitions between different sections, stores, etc. Hence, profiling and identifying group and individual behaviors, can aid in more accurate surveillance.

Social Event Detection: As shown in Jayarajah et. al. [6], the social ties amongst a group or crowd of people gathering at a place are useful features in detecting transient social events, both indoors (at campus scale) and outdoors (at city scale). Group detection is a key ingredient in deriving such trajectory-based ties from longitudinal observations, constructing physical social networks, and extracting network properties. It is sufficient to consider deviations in the volume of people present (or absent) to detect *high intensity* events. But, by detecting that groups of strongly connected individuals (*friends*) are gathering at certain locations, it is also possible to detect low intensity (e.g., friends/families attending a theatrical performance) events that affect only a small fraction of people.

3. SYSTEM OVERVIEW

We first describe our multi-modal group detection system in detail. The key intuition here is that people in the same group would exhibit correlated or similar mobility patterns. The system consists of three main blocks: WiFi-based group detection, video-based group detection and probabilistic group membership classification, and relies solely on server-side WiFi-based localization and video feeds (See Figure 1).

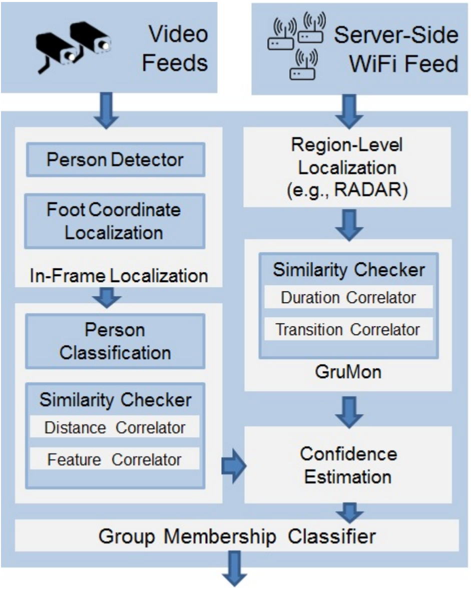

Figure 1: System Architecture.

[a] WiFi-based Group Detection: From the server-side WiFi traces, RADAR [1] (or any similar technique) is used to infer coarse-grained indoor location (see Figure 2), at a *section* level (typically spanning 6-8 meters) of devices carried by persons. Then, a traveling companion-like detection mechanism (presented in *GruMon* [15]) is used on location traces to detect groups. We consider two or more persons to be a group if they *transition* between two stay regions (a 'stay region' is a section where a mobile device is seen to be stationary for a period $\geq X$) together within a detection window T. Additionally, to accommodate the phenomenon of updates with variable delay, we allow for a period of lag (T_l) after T, within which a set of people can still be considered a group, but with reduced confidence (based on the *duration of the time spent co-located*).

[b] Video-based Group Detection: From a continuous video stream surveying the connecting corridor between two stay regions (see Figure 2), for every detection window T, we use frame-wise distance and feature similarity based heuristics (computed over the video frames) to detect groups. We describe in detail in Section 3.1.

[c] Group Membership Classification: Based on the two independent group detection system outputs, a combined result is computed based on the cardinality of detected groups and their confidence levels. We describe the details of the fusion algorithm in Section 4.

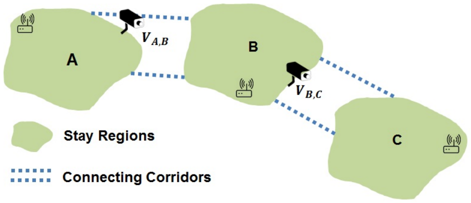

Figure 2: Our Approach: Video ($V_{A,B}$, $V_{B,C}$) captures transitions between WiFi-based stay regions (A & B and B & C, respectively).

3.1 Video-based Group Detection

Video-based detection of groups involves three main steps: (1) detecting and localizing a person in a frame, (2) second, from consecutive frames, classifying and tracking persons, and (3) lastly, based on trajectories of persons across frames, detecting groups.

3.1.1 In-Frame Localization

Person Detection: We analyze each video frame to detect the presence of "persons". We investigate different person-detectors – HAAR-cascades [18], HOG+SVM [2], DPM [16] and YoLo [14] – , and chose YoLo as it provided the best performance (i) under medium-to-bright indoor lighting conditions, (ii) with occlusion (detects even when certain parts of the body are not visible in the frame), and (iii) irrespective of the direction of the person walking (e.g., HAAR-based classifiers depend on the facial features, which cannot be obtained if the person is walking away from the camera).

From the *bounding boxes* of persons detected, we infer the coordinate of the foot as $< (x_{min} + x_{max})/2, y_{min} >$ where the bounding box is defined by $< x_{min}, y_{min}, x_{max}, y_{max} >$.

Distance-Corrected Localization: We note that the perceived distance between two persons, or the perceived distance a person has moved between frames, is a function of both the actual distance and the distance of the person from the camera – for e.g., the perceived distance between two people is larger when they are closer to the camera as opposed to when they are far, although the real distance between the two may remain the same. To correct for this, during calibration, we compute a perspective transform matrix by (1) first choosing four bounding coordinates along the edges of the *vanishing* corridor visible in the frame, and (2) then identifying the coordinates of the *corrected* corridor. We apply this transformation on the raw foot coordinate to compute a distance-corrected "real world" coordinate.

3.1.2 Person Classification and Tracking

To match two person objects detected in two consecutive frames as the same person (or not), we use both distance-based and feature similarity based measures. Empirically, we found that at an average walking speed, and a frame rate of 30 fps, a person moves less than 50 pixels between frames. However, it is possible for two persons walking together to be each mistaken for the other when viewed over quick consecutive frames. To avoid this, we additionally obtain the SIFT-features[8] and compute the pairwise feature similarity score as $\frac{numberofgoodkeypointsmatched}{numberofkeypointsmatched}$ between the bounding boxes. We consider $goodkeypoints$ as those keypoints that have a distance ratio better than 0.7 as described in Lowe et al. [8]. We compare consecutive frames and declare that a detected person matches a previously detected person using a distance threshold of 50 pixels and a feature similarity score threshold of 0.3. We decided on these threshold values through trial and error.

3.1.3 Group Detection

We track persons over a block of frames of size F_B, and consider both (1) the number of frames any two persons were seen together, and (2) the average distance between the trajectories of the two persons across the co-located frames. If they are seen together for more than F_G frames, and the average distance is less than T_d, then they are declared as a group. If A and B are a group, and B and C are a group, then by transitivity, we consider A, B and C as a group. We evaluate the choices of F_B, F_G and T_d in Section 5.

4. FUSION OF WIFI-BASED AND VIDEO-BASED GROUP DETECTION

We present our probabilistic group membership algorithm in Algorithm 1. Besides the video-based detector, the WiFi-based detector outputs groups every $T secs$. We allow a lag of $T_l secs$, which is a multiple of T, such that devices that same exhibit the same transition, but offset by less than T_l secs, are considered as *potential* groups. We define the confidence of a detected candidate group as the proportion of time the members of the group have spent together (prior to the common transition within $T + T_l$) of the maximum time any subset of the group had spent together. For example, if A and B transition together T, and C transitions within the next T_l, then the confidence in (A, B, C) being a group is less than that of (A, B) being in a group. We formulate this confidence, P_S^W, as in Equation 1 where S is the member set, S' is every proper subset of S and t_S is the overlapping time duration of S.

Similarly, we define the confidence in a group detected by video, P_S^V, as in Equation 2. Here, d_S is the average distance (in pixels) between pairs in S and T_d is the Distance Threshold within which any two detected persons are considered to be in a group. Additionally, for video-based groups, we consider the temporal distance between any two groups for declaring sub-groups. We define the time at which a group was last detected by video as t_S^V. Then the confidence that two groups S_1 and S_2 are separate groups, P_{S_1,S_2}^T, is defined as in Equation 3 where T_S is a Temporal Distance Threshold.

Subgroup Detection: If $P_{S_1,S_2}^T \geq \alpha$, then the likelihood that S_1 and S_2 are separate groups is high – although a single WiFi group is detected for the same T with some P_S^W.

$$P_S^W = \frac{t_S}{max(t_{S'})} \tag{1}$$

$$P_S^V = 1 - \frac{d_S}{T_d} \tag{2}$$

$$P_{S_1,S_2}^T = 1 - \frac{abs(t_{S_1}^V - t_{S_2}^V)}{T_S} \tag{3}$$

Hidden Nodes: For the same decision window, if $N_W < N_V$, we declare that there were $N_V - N_W$ hidden (an unseen WiFi device) nodes. N_W and N_V are the cardinality of the groups detected using WiFi and video, respectively, over the $(t, t + T + T_l)$ period. For this initial work, we do not consider the rare case $N_V > N_W$, i.e., when the video analytics exhibits false negatives, over an *entire block of F_B frames.*

Location Lags: For the same decision window, if $N_W = N_V$, and $P_S^V > \beta$ although $P_S^W < \theta$, then we declare that S was indeed a group and the decrease in P_S^W is likely due to delays in localization. Here, α, β and θ are appropriately chosen thresholds on the probabilities for group declaration.

Assumptions: This methodology assumes that (i) the location system does not generate erroneous location estimates, and that a location report for a 'stay point' is always generated within T_l secs of the true transition time, and (ii) multiple groups do not move identically during the same T.

We evaluate the cases of *hidden nodes* and *subgroups* in Section 5.

5. EVALUATION

We first evaluate the two group detection modalities in isolation in order to understand the impact of the parameter choices on accuracy. We then summarize early results from the combined group

Algorithm 1 Probabilistic Group Membership

```
Input: region level location, in-frame location
Output: groups detected, confidence
  while $t \geq (2T + T_I)$ and $t$ mod $T \equiv 0$ do
    $S_W, P_S^W, N_W = detectWiFiGroups(t, t + T + T_I)$
    $S_V, P_S^V, N_V = detectVideoGroups(t, t + T + T_I)$
    if $N_W \equiv N_V$ then
      $initializeSubGroupDetection(S_V)$;
    else if $N_W < N_V$ then
      $declareHidden(N_V - N_W)$;
      $initializeSubGroupDetection(S_V)$;
    else
      % Deferred as future work
    end if
  end while
```

detection system in overcoming the challenges of hidden nodes and subgroups.

5.1 Study Details

We conducted an instrumented micro-study with 10 participants from our lab. The participants were split into 5 groups (or individuals, see Table 1) and were instructed to spend time at two stay regions and alternate (or *transition*) between them at scheduled times. The corridor (roughly $3m \times 1m$ area) was monitored by a standard web-cam at a resolution of 480×640, and a frame rate of 30 fps. The participants carried instrumented Android phones that scanned for WiFi devices and recorded their RSSI values every 100 ms. The 10 participants were asked to transition back and forth four times, spending at least 5 minutes at each stay region, resulting in a total of 20 group (& individual) episodes.

To emulate *hidden nodes*, WiFi traces of 3 participants chosen at random were ignored during WiFi-based group detection. Further, to emulate *sub-groups*, 3 of the 5 sets were asked to move around with only a few seconds apart. The details of the group configurations are given in Table 1.

Group	Members	Scenario
Group A	2	ideal case with all nodes known and clear separation from other groups
Group B	2	followed A after 5 mins
Subgroup C	2	followed B after 1 min
Subgroup D	3	followed group C after 2 secs.
Subgroup E	1	followed group D after 2 secs

Table 1: Group configurations of the study.

5.1.1 Accuracy Metrics

We report precision and recall values in the following way: (1) $recall = \frac{N_{detected}}{N_{episodes}}$ and (2) $precision = \frac{N_{correct}}{N_{detected}}$. Here, $N_{detected}$ is the number of episodes that were detected whereas $N_{episodes} = 20$, is the total number of episodes. In computing precision, we allow for partial matches – i.e., the contribution of a partially detected group towards precision as ($size_{detected}$ is the size of the detected group and $size_{actual}$ is the size of the actual group):

$$1 - \frac{size_{detected} - size_{actual}}{max(size_{detected}, size_{actual})} \quad (4)$$

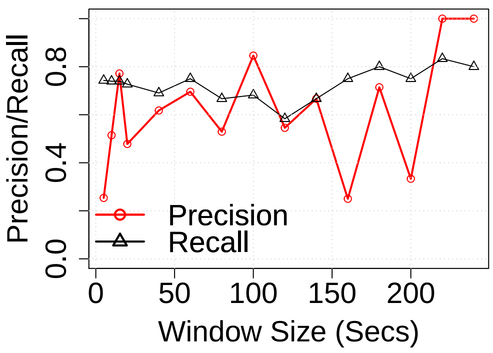

Figure 3: Accuracy vs. Observation Window Size Trade-Off in Server-Side WiFi-based Detection.

5.2 Choosing Window Size T for WiFi-based Detection

To identify an appropriate choice for T, we used server-side (with an update rate of roughly 5 secs) localization traces of three participants who spent an hour together transitioning between four stay regions. In Figure 3, we plot the precision and recall (y-axis) for different window sizes (x-axis). We observe that as the window size increases, the detection precision improves gradually. We have shown previously in Sen et. al. [15], that with client-side localization and higher latencies ($\approx$ 10 mins), the recall can increase up to 92%. For the purpose of our study, we choose $T = 60$ secs as a good trade-off between accuracy and the detection latency.

5.3 Choosing Parameters for Video-based Detection

We varied the three parameters, (1) *Block Size (F_B)* which is the number of frames processed before a group is detected, (2) the *Group Candidate Threshold (F_G)* which is the number of frames within the block any two persons have to be seen to be considered a candidate for group, and (3) the *Distance Threshold (T_d)* which is the maximum distance (in pixels) between any two persons to be considered a candidate for group, averaged over frames they were co-located within a block.

In Figure 4 and Figure 5, we plot the precision/recall (y-axis), for $F_B \in [15, 30, 60, 75]$ and $F_G \in [5, 10, 15]$ with $T_d = 200$ in all cases. We observe that the precision is robust with varying block sizes. However, we note that the recall drops significantly with both increase in block size and frames to compare. In the case of $F_B \geq 60$, i.e., a block duration greater than 2$seconds$, the ability to differentiate between the subgroups C, D, E drops causing the recall to drop. We also varied the Distance Threshold (T_d) which is the upper limit on the average distance (in pixels) within which two persons are considered to be in a group. Further, for the same block size, requiring more than 5 frames to compare causes 20% less groups to be detected (0.2 drop in recall). For values 150, 200 and 250, we did not observe any significance difference in the precision or recall. Overall, we found that $F_B = 15$, $F_G = 5$ and $T_d = 250$ as the combination with the best performance.

	SG Only (WiFi)	SG + HN (WiFi)	Video Only	Fusion
Precision	0.58	0.5	0.81	0.81
Recall	0.6	0.6	0.8	0.9

Table 2: Overall Accuracy under Perfect Localization. HN-Hidden Nodes, SG-Subgroup.

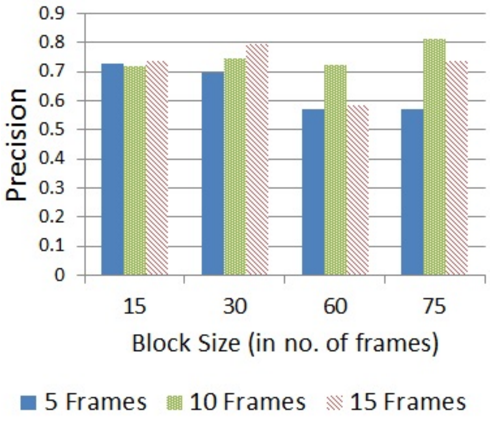

Figure 4: Precision vs. Block Sizes (F_B) and Group Candidate Thresholds (F_G).

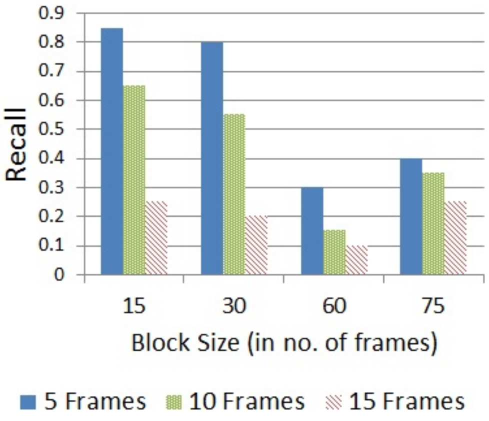

Figure 5: Recall vs. Block Sizes (F_B) and Group Candidate Thresholds (F_G).

5.4 Group Detection Under Perfect Localization

For the purpose of this study, we evaluate the system under perfect location conditions – i.e., we assume the location provided by WiFi is both accurate and does not incur any delays. Hence, we simulate the indoor location traces based on ground truth locations and times. We use $F_B = 15$, $F_G = 5$, $T_d = 250$, $T = 60$ secs, $T_I = 60$ secs, $T_S = 600$, and $\alpha = 0.01$. As the block size is much smaller than the time for which a group is seen by the camera, the group will be detected multiple times over consecutive frames – for e.g., members of the same group enter and exit one by one into the camera's field of view. To avoid counting them as separate group instances, we look at the time separation between the detection times, if it is less than a threshold of 0.5 secs, we consider them as the same group. This improved the precision of video-based detection by ≈ 0.1.

We report our results in Table 2. Overall, we observe a precision of 0.81 and recall of 0.9 for the study with the 20 group episodes. In the case of *Subgroups only*, we use location traces of all participants with C, D, E transitioning as different groups, but very close temporally (i.e., 40% of the groups were subgroups). In the *Subgroups and Hidden Nodes* case, the traces of one participant each from groups B, C, and D were removed (i.e., 33% of the participants were hidden). We observe that video detects *all* subgroups and 67% of the hidden nodes resulting in a 30% improvement in precision, and a 10% improvement in recall over the WiFi-only approach. WiFi detected at least two group instances which the video missed due to false negatives in person detection. Hence, considering the combined output of the fusion algorithm, the recall increases by another 10% to 0.9.

6. DISCUSSION AND FUTURE WORK

Current Limitations: As we noted earlier, video-based person detection suffers from both false-positives and false-negatives (See Figure 6 and Figure 7).

Handling False Positives: We consistently observe that the falsely detected bounding boxes overlap significantly with adjacent bounding boxes of correctly detected persons. To handle this, we propose to consider the area of overlap and if it is above a certain threshold, to consider them as a single, merged detected person.

Handling False Negatives: In the second case, we noticed that a person is often detected in a few consecutive frames, missed in the next few (due to various reasons including lighting & occlusion) and then is detected as a *new* person because the distance and feature similarity thresholds exceed since the last detection. To deal with this, we plan to *interpolate* the trajectory of the person in the missed frames, using feature matching, until the same person is detected again.

Ongoing Work: In this paper, we presented preliminary results under ideal localization. In practical scenarios, we expect the location traces to be noisy. Currently, we are extending Algorithm 1 to account for location errors (accuracy and latency issues). Further, in our limited trials, none of the groups occurred together within the same frame, or involved participants who changed their movement direction mid-way through the observation period – in future, we intend to expand our study with more real-world group scenarios. Moreover, we must extend on our approach to a larger, indoor space where we have multiple cameras (surveying transition regions) connecting multiple stay points.

7. RELATED WORK

WiFi-based Group Detection: Both Liu. et. al. [7] and Sen et. al. [15] describe group detection using server-side WiFi- based indoor location traces. However, the accuracy of the system drops due to the presence of hidden nodes and groups that transition within short inter-group intervals. We use a location-only version of *GruMon* [15] for the pure-WiFi based group detection in this work.

WiFi/Location+Video Sensor Fusion: Several works [4, 17, 12] in the past have described the opportunities and challenges in fusing WiFi with video data for tracking people. In Nandakumar et. al. [12], the authors share their initial thoughts on how the combination of vision tracking (that suffers from fragmentation, occlusion and spurious objects) and sporadic WiFi signals can be fused for accurate tracking of users in-store. The E-V system described in Zhu et. al. [17] locates a person based on his/her E and V signals, i.e., his/her electronic footprint and and his visual appearance in the video, respectively. Similarly, Jamtgaard et. al. [4] describe a technique for tracking targets based on spatio-temporal correlation in WiFi and video. In contrast, Miyaki et. al. [11] propose a particle filtering approach for continuous target tracking using the two modes. Recently, in Ganti et. al. [3] the authors describe an orthogonal problem of *entity reconciliation* – i.e., to detect an entity (e.g., person) across several sources of mobility data such as smartphone-generated GPS traces and videos from multi-camera networks, to provide a combined, global view of the entity. Our work is different (albeit complementary) in that we do not attempt to track an individual or match persons/targets across the multiple modes.

Accurate Indoor Localization: Although previous literature have described techniques [9, 13] that can offer much finer accura-

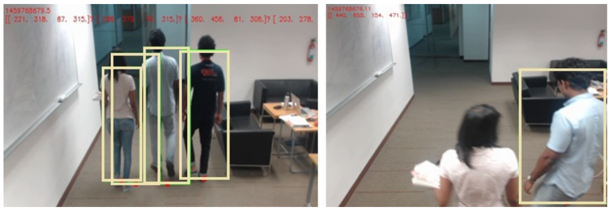

Figure 6: False-Positives and False-Negatives in Video-based Person Detection. The bounding boxes indicate detected person bounding boxes. Left: an extra person is detected falsely sharing high overlap with two correctly detected persons, Right: a person close to the camera is not detected due to blur.

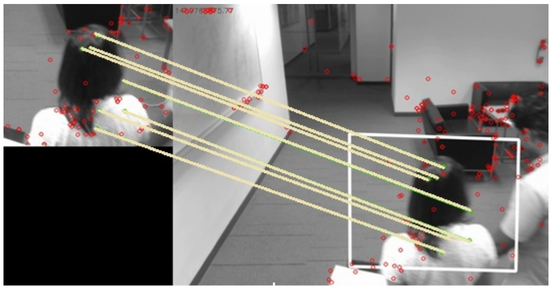

Figure 7: Correcting for a False Negative using Feature Matching. Left: the bounding box of a person detected in the previous frame, Right: detecting the same (missed) person by searching the following frame for matching SIFT-features.

cies (e.g., sub-meter scale), they are not as pervasive as server-side, Wi-Fi based techniques because they either require large-scale deployment of specialized hardware (e.g., [13]), or client-side support (e.g., dead reckoning on the smartphones [9]). Our work describes a fusion framework that leverages *practical* indoor localization that inherently suffers from accuracy and latency issues.

8. CONCLUSION

In this paper, we described a group detection pipeline for video and proposed a novel system which fuses both outputs from WiFi-based and video-based group detection. By taking into account both levels of confidence in detection of the independent sources of mobility and carefully chosen spatial and temporal thresholds, we show that the group detection accuracy can be improved by at least 20% over the traditional WiFi-only based solutions. Although our study is preliminary, and further investigation over a larger scale is warranted, we believe that this works opens up discussion on fusing multi-modal mobility information for high accuracy, low latency, mobility-aware futuristic systems.

9. ACKNOWLEDGEMENTS

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its IDM Futures Funding Initiative. Kasthuri Jayarajah is supported by an A*STAR Graduate Scholarship. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the granting agencies or Singapore Management University.

10. REFERENCES

[1] P. Bahl and V. N. Padmanabhan, *Radar: an in-building rf-based user location and tracking system*, IEEE., 2000.

[2] N. Dalal and B. Triggs, *Histograms of oriented gradients for human detection*, Proc. Of CVPR'05.

[3] R. K. Ganti, M. Srivatsa, and B. S. Manjunath, *Entity reconciliation in a multi-camera network*, Proc. of ICDCN'16.

[4] M. Jamtgaard and N. Mueller, *Target localization utilizing wireless and camera sensor fusion*, December 24 2013, US Patent 8,615,254.

[5] K. Jayarajah, Y. Lee, A. Misra, and R. K. Balan, *Need accurate user behaviour?: Pay attention to groups!*, Proc. of UbiComp '15.

[6] K. Jayarajah, A. Misra, X. W. Ruan, and E. P. Lim, *Event detection: Exploiting socio-physical interactions in physical spaces*, Proc. of ASONAM'15.

[7] S. Liu, S. Wang, K. Jayarajah, A. Misra, and R. Krishnan, *Todmis: Mining communities from trajectories*, Proc. of CIKM'13.

[8] D. G. Lowe, *Distinctive image features from scale-invariant keypoints*, Int. J. Comput. Vision **60**.

[9] Alex T Mariakakis, Souvik Sen, Jeongkeun Lee, and Kyu-Han Kim, *Sail: Single access point-based indoor localization*, Proc. of MobiSys'14.

[10] S. J. McKenna, S. Jabri, Z. Duric, A. Rosenfeld, and H. Wechsler, *Tracking groups of people*, Computer Vision and Image Understanding **80** (2000), no. 1.

[11] T. Miyaki, T. Yamasaki, and K. Aizawa, *Tracking persons using particle filter fusing visual and wi-fi localizations for widely distributed camera*, Proc. of ICIP'07.

[12] Rajalakshmi N., Swati R., Krishna C., Venkata N. ., Lili Q., Aishwarya G., Saikat G., Deepanker A., and Aakash G., *Physical analytics: A new frontier for (indoor) location research*, MSR-TR-2013-107, 2013.

[13] Nissanka Bodhi Priyantha, *The cricket indoor location system*, Ph.D. thesis, Massachusetts Institute of Technology, 2005.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, *You only look once: Unified, real-time object detection*, arXiv preprint arXiv:1506.02640 (2015).

[15] R. Sen, Y. Lee, K. Jayarajah, A. Misra, and R. K. Balan, *Grumon: Fast and accurate group monitoring for heterogeneous urban spaces*, Proc. of SenSys '14.

[16] S. Tang, M. Andriluka, and B. Schiele, *Detection and tracking of occluded people*, Int. J. Comput. Vision **110** (2014), no. 1, 58–69.

[17] J. Teng, J. Zhu, Boying Z.g, D. Xuan, and Y. F. Zheng, *E-v: Efficient visual surveillance with electronic footprints*, Proc. of INFOCOM'12.

[18] P. Viola and M. Jones, *Rapid object detection using a boosted cascade of simple features*, Proc. Of CVPR'01.