Poster Abstract: WiFi Sensors Meet Visual Tracking For An Accurate Positioning System

Savvas Papaioannou, Hongkai Wen, Zhuoling Xiao, Andrew Markham, and Niki Trigoni
Department of Computer Science, University of Oxford. Email: firstname.lastname@cs.ox.ac.uk.

***Abstract*—In this poster abstract, we propose a new positioning technique that can localize people by combining WiFi information from their mobile devices with visual tracking. We show that the proposed approach can improve visual tracking by resolving motion and appearance ambiguities while at the same time can uniquely identify each person with their device ID.**

I. Introduction

The development of practical and accurate indoor positioning systems has received a lot of attention from the research community. One commonly used positioning technique is vision-based tracking. Many existing vision-based tracking algorithms are capable of accurately estimating the trajectories of multiple targets from video footage. However, the estimated trajectories are usually *anonymous*. Visual identification (e.g. face recognition) cannot always be applied since it requires knowledge on the mapping between IDs and pictures, and can be computationally expensive and privacy intrusive. On the other hand, the WiFi-based positioning systems are able to localize the mobile devices carried by the targets. Therefore, the WiFi systems possess the knowledge of the targets' ID, but typically require a stable radio map, use many APs and often a floor plan to achieve good accuracy.

Motivated by the above problems, we propose a new positioning technique which is able to perform accurate localization and privacy-preserving identification. The key idea is to exploit the existing WiFi and camera infrastructure, which is available in most of today's large indoor environments. We use visual tracking techniques to detect moving objects in the camera footage, and generate anonymous tracklets based on a motion model. The tracklets are then fused with WiFi signal strength measurements to produce accurate trajectories of each target. Unlike most of the existing work which uses WiFi information to *match* the trajectories produced by visual tracking, our approach incorporates WiFi measurements in visual tracking to *generate* the trajectories, and thus can achieve similar performance with less infrastructure and noisy signals.

II. Proposed Approach

A. System Architecture

The proposed approach contains two components: a *foreground detector* and a *tracker*, as shown in Figure 1. We assume the indoor environment is covered by one calibrated stationary camera, and multiple WiFi access points (APs) with known locations are deployed. A number of people (targets) are moving around with their mobile devices

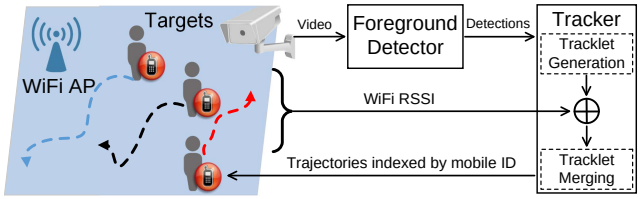


Fig. 1. The architecture of the proposed system.

Fig. 2. The detections generated by our implementation of foreground detector.

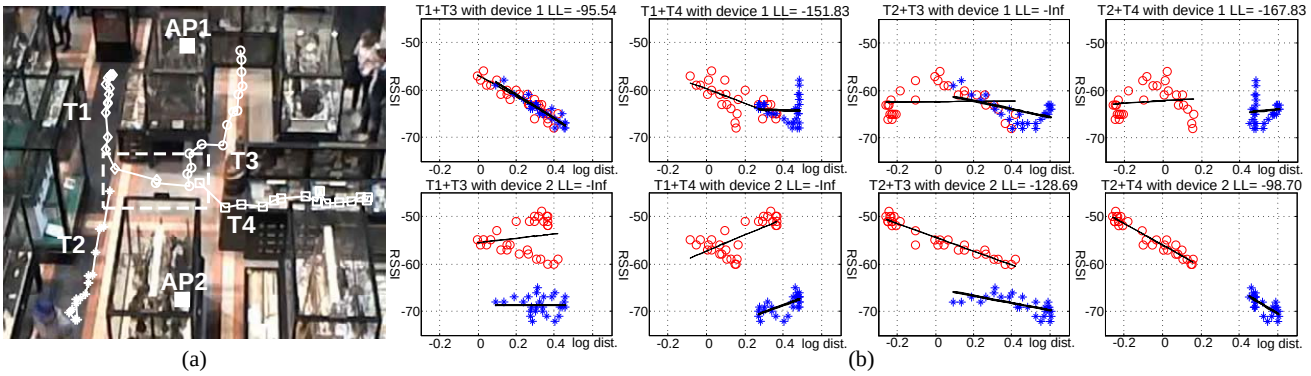
receiving WiFi beacons from the APs periodically. The captured video stream is processed by the foreground detector to extract detections of moving objects, and then the tracker fuses the visual detections with WiFi RSSI measurements collected by the mobile devices to a) generate accurate trajectories, and b) identify the target ID of each trajectory.

B. Foreground Detector

The foreground detector detects any moving objects in the camera footage, and can be implemented in a number of different ways [1]. Comparing to the more sophisticated object detection techniques (e.g. [2]), foreground detector is lightweight, and it can be used in real-time in embedded camera networks [3]. However, the detections generated by a foreground detector can be very *noisy*. We have observed three distinct cases in our preliminary experiments where this happens (as shown in Figure 2): a) multiple detections are generated for one moving object (D1 and D2), b) a detection contains no moving object at all (D4), and c) one detection contains multiple moving objects (D5).

C. WiFi-based Tracker

The key component of the proposed approach is the WiFi-based tracker, which works in two steps: 1) *tracklet generation* and 2) *tracklet merging*. Concretely, suppose we have m targets. The input video is divided into s segments, each of which has n frames. The tracker first generates short trajectories (tracklets) for the s video segments, and



(a) (b)

Fig. 3. (a) Tracklets generated for two targets in two video segments, where $T1$, $T2$ are in the first segment, and $T3$, $T4$ in the second. (b) WiFi signal strength measurements can help find the correct solutions: $T1$+$T3$ with device 1 and $T2$+$T4$ with device 2 (the first plot and the last plot).

then merges them to produce the final trajectories of the m targets.

Tracklet generation: For each segment a weighted directed acyclic graph $G = (V, E, W)$ is generated, where the vertices V represents the detections computed by the foreground detector, and are partitioned into n disjoint groups according to frames. The edges E are defined between the vertices in different groups to form a trellis diagram. The weight w of an edge e describes the appearance and motion costs that two detections belong to the same trajectory. The appearance cost is determined by the color histogram intersection between the two detections. The motion cost is calculated by comparing the distance between the two detections with a motion model learned from the data, which constrains the movement speed. For one target, we use a Viterbi-like technique to traverse the trellis graph to find the tracklet with the minimum cost by selecting one detection from each group of vertices. We then remove the found tracklet from the graph and perform the same procedure for the next target until all m tracklets are found.

Tracklet merging: In this step, the tracker merges the tracklets generated for each segment to produce the complete trajectories, and assigns the correct target IDs to them. We use a sliding window algorithm, which considers the tracklets in k segments at a time. Concretely, the algorithm performs k-partite matching to find the most likely trajectory of a target. It iteratively searches the space of all possible solutions that contain k tracklets within the current window, guided by a cost function: $C(l) = w_A C_A(l) + w_M C_M(l) + w_W C_W(l)$, where l is a solution (i.e. a possible trajectory associated with a target ID), C_A, C_M, C_W are the cost functions for appearance, motion and WiFi respectively, and w_A, w_M, w_W are the normalizing weights. $C_A(l)$ is computed from the pairwise intersections of the average color histograms of tracklets in l. $C_M(l)$ is evaluated by comparing the pairwise distances between the tracklets in l (begin and end points) with the motion model. The WiFi cost $C_W(l)$ describes how consistent is the solution l with the WiFi RSSI measurements of a device (carried by a target). Given the locations of the APs, the log-distances between the points on the solution l and the APs can be known exactly. If l is correct, then the relationship between the log-distances and the sequence of signal strength measurements should be *linear*, governed by the radio propagation model. With this intuition, our tracker performs Bayesian linear regression on the log-distances and the RSSI measurements, and $C_W(l)$ is defined as the log-likelihood that the observed RSSI sequence agrees with the trajectory l under the radio model. Figure 3 shows how WiFi can help to associate the tracklets correctly. We consider two targets (two different devices) in two video segments, where ($T1$,$T2$) and ($T3$,$T4$) are the tracklets in the 1st and 2nd segment respectively. In this case it is very difficult to find the correct solution based only on motion (assuming similar appearance) since the targets are close in the highlighted region. Figure 3(b) shows the relationship between the log-distances and the observed RSSI measurements from 2 APs for all possible solutions (2 devices $\times$ 4 possible trajectories), where only the two solutions: a) $T1$+$T3$ with device 1 and b) $T2$+$T4$ with device 2 are correct (also indicated by the higher log-likelihood (LL)). Note that the same technique can also be used to deal with occlusions (suppose no trajectory was in the highlighted box in Fig. 3(a)), since WiFi measurements can correctly connect the tracklets to fill the gap. Our initial results in a real setting (a museum) indicate that the proposed technique can be used to resolve motion and appearance ambiguities.

III. Acknowledgment

We would like to thank Laing O'Rourke for funding this research and also the Pitt Rivers Museum for allowing us to conduct our experiments in the museum's space.

IV. Conclusion

We propose a new positioning system that integrates WiFi information with visual tracking. The novelty of our approach is that it leverages WiFi measurements to improve the performance of visual tracking, and offer accurate localization and identification at the same time.

References

[1] C. Stauffer and W. E. L. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. CVPR*, 1999.

[2] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *PAMI*, vol. **32**, no. 9, pp. 1627–1645, 2010.

[3] Y. Shen, W. Hu, J. Liu, M. Yang, B. Wei, and C. T. Chou, "Efficient background subtraction for real-time tracking in embedded camera networks," in *Proc. SenSys*, 2012.