RGB-W: When Vision Meets Wireless

Alexandre Alahi Albert Haque Li Fei-Fei
Computer Science Department, Stanford University
{alahi,ahaque,feifeili}@cs.stanford.edu

Abstract

Inspired by the recent success of RGB-D cameras, we propose the enrichment of RGB data with an additional "quasi-free" modality, namely, the wireless signal emitted by individuals' cell phones, referred to as RGB-W. The received signal strength acts as a rough proxy for depth and a reliable cue on a person's identity. Although the measured signals are noisy, we demonstrate that the combination of visual and wireless data significantly improves the localization accuracy. We introduce a novel image-driven representation of wireless data which embeds all received signals onto a single image. We then evaluate the ability of this additional data to (i) locate persons within a sparsity-driven framework and to (ii) track individuals with a new confidence measure on the data association problem. Our solution outperforms existing localization methods. It can be applied to the millions of currently installed RGB cameras to better analyze human behavior and offer the next generation of high-accuracy location-based services.

1. Introduction

The analysis of human behavior in indoor spaces significantly improved over the recent years as a result of complementing RGB data with the depth modality (RGB-D) [34, 18, 7, 2]. However, these setups are rare and often too costly to deploy. Today, millions of spaces are monitored by a single RGB camera. The challenges with these monocular views lie in the depth estimation and self-occlusion problems. To address these challenges, we propose to complement RGB data with an additional "quasi-free" modality, namely wireless signals (e.g. wifi or Bluetooth) emitted by cell phones, referred to as RGB-W (see Figure 1). Recent studies have shown that over 50% of visitors in public spaces leave their wifi enabled, and several municipal governments are planning large-scale wifi implementations. Meanwhile, beacon technology (Bluetooth Low Energy) is also being deployed in public spaces to enable location-based services such as self-guided tours, item delivery, or to perform role-based activity understanding in hospitals. To benefit from these services, visitors and staff intentionally agree to share their wireless signal.

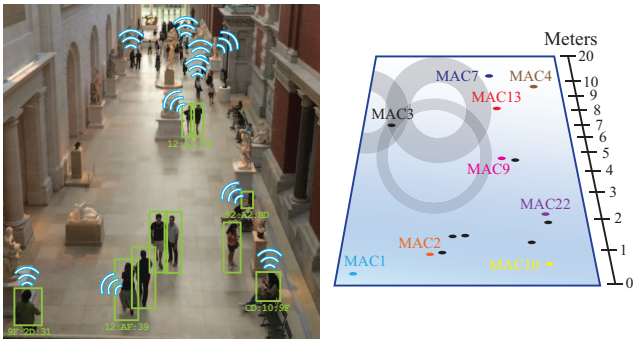


Figure 1: Illustration of a scene captured with RGB-W data. The **W** data represents the received signal strength (RSS) from individuals' cell phones (through wifi or Bluetooth) with their corresponding unique identifier (e.g., MAC address). We aim to jointly locate and track individuals with our proposed ring-based image representation of the wireless signals.

In this paper, we aim to improve the localization and tracking of individuals with RGB-W data. This has numerous benefits for applications ranging from space analytics for safety, security, and behavioral studies, to location-based services using smartphones. For the sake of clarity, we refer to wifi, Bluetooth, or beacon signals, as **W** data throughout this paper. **W** data provides a stream of packets from each phone describing the received signal strength (RSS) of the packets and their origin – a unique identifier commonly called a mac ID. The tuple {RSS, mac ID} is captured through the **W** modality and serves as an additional source of information to better solve vision tasks with a RGB camera.

The underlying motivation behind RGB-W data is the complementary nature of the two modalities. On one hand, RGB-based methods can accurately locate and track individuals in the absence of occlusion, but in crowded scenes, their performance deteriorates. On the other hand, **W** data does not suffer from the occlusion problem and can solve the data association across time with the observed mac ID, but cannot precisely locate in 3D. To fuse the advantages

of both modalities, the following challenges need to be addressed with RGB-W data:

1. Noisy **W** - The RSS is highly dependent on the environment and signal interference, exhibiting variances of 10 dBm (i.e., localization errors of several meters).
2. Sparse **W** - Only a subset of individuals present in a scene may broadcast **W** signals. Additionally, the temporal sampling rate of **W** data is lower than the RGB frame-rate (e.g., 2-5 **W** samples per second)
3. Incomplete RGB - The RGB streams lack depth information and experience strong occlusion issues.

We aim to address the above challenges by jointly processing RGB-W data in a unified framework. Our contributions are as follows: (i) we suggest a new image-driven representation of the **W** data to enable joint reasoning with RGB images (Section 3). We represent the **W** data as an image which embeds the estimated radius as well as its estimated variance (error bounds) to fully model the information available from the **W** data. (ii) We present a sparsity driven framework with a cascade of ℓ_1 solvers to locate individuals with RGB-W data (Section 4). We fuse both foreground and ring images into a single dictionary to jointly solve the ground occupancy of individuals. (iii) We demonstrate the impact of RGB-W on the tracking framework by solving the mac assignment task given noisy observation of the data. (iv) Finally, we share our RGB-W dataset as well as the data collection protocol to ease future work.

The paper is structured as follows: First, we briefly present existing localization and tracking efforts that use RGB and **W** signals in Section 2. Then, we describe our image representation of **W** , followed by the RGB-W based localization and tracking framework 5. We conclude by presenting quantitative results with respect to previous RGB and **W** based localization and tracking techniques.

2. Related Work

Locating and tracking individuals has piqued the interest of various communities ranging from computer vision to sensor networks. We review localization methods using RGB only, **W** only, and attempts using both modalities.

RGB-Based Localization and Tracking Pedestrian detection can be achieved using a single image and image classification techniques such as R-CNNs or deformable parts models [13, 16, 35, 4]. Individuals are detected in the image plane as opposed to 3D coordinates of people in the real world. With a calibrated camera, the authors in [11, 1] have shown that it is possible to map a detected bounding box to the real world coordinates.

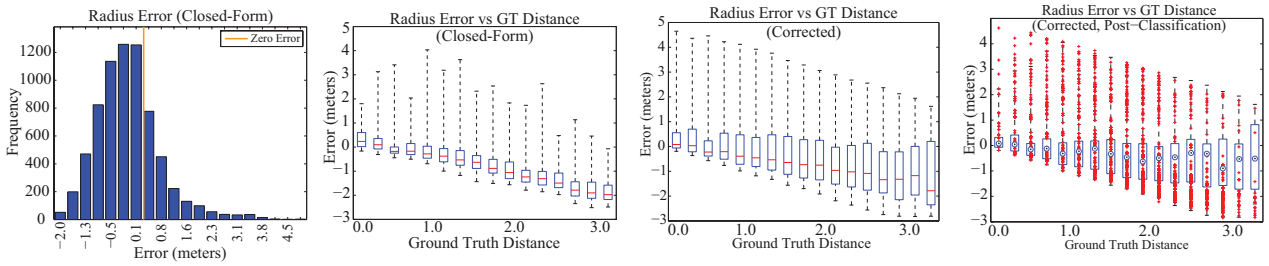
Algorithms with high levels of confidence have been proposed to locate crowded people with a single top view or several head-level overlapping field-of-views [10, 12, 21]. In [21], Khan and Shah locate people on the ground where decent foreground silhouettes are observed in several camera views. Alahi *et al*. in [1] proposed a sparsity driven framework to handle noisy observations given a precomputed dictionary. Recently, Golbabaee *et al*. presented a model in [17] for detecting and tracking people in real-time. Instead of solving a convex relaxation of the detection step with iterative shrinkage, they proposed a greedy algorithm inspired by the set cover problem.

Once individuals are located on the ground, various graph-based algorithms can be utilized to track them. Recently, global optimization was performed with linear programming to address the data association problem [5, 24]. It outperforms previous works based on Markov Chain Monte Carlo [22] or inference in Bayesian networks [31]. The data association problem is expressed as a graph theoretic problem for finding the best path for each point across frames. RGB-W data provides a unique identifier for individuals sharing their **W** signals, and thus, can be naturally integrated into such formulations.

W-Based Localization Several studies aim to leverage wifi or Bluetooth to perform localization, especially in indoor environments [37, 19]. Attempts at **W** based localization can be categorized into two groups: (i) fingerprint databases and (ii) trilateration using signal propagation models. Fingerprint databases store signal strengths at various known reference points. Each reference point contains a unique fingerprint of signal strengths. The fingerprint database emulates a lookup table during real time localization. Distance based methods are typically employed to find the nearest reference point [29]. Because reference points must be manually collected offline, fingerprint databases are often time consuming and expensive.

Signal trilateration and propagation models have been well documented and are able to estimate position within 2 meters [30, 27]. To estimate distance from each antenna, a variety of models can be employed: Gaussian models [14], Monte Carlo [6], Bayesian [25], Hidden Markov Models [23], and radio propagation models [3] have been presented over the years. Ring overlapping approaches have been proposed in the past [26, 36], however these methods tend to require additional calibration, both at the antenna and the phone level.

RGB-W-Based Tracking There have been past attempts to fuse both the RGB and **W** modalities. In [28], Miyaki *et al*. used a particle filter approach to perform outdoor tracking using a distributed camera setup and a RSS centroid approach similar to [8]. Although Miyaki *et al*. were successful at outdoor tracking, the use of a GPS to collect the ground truth and error of up to 18 meters makes their approach unfeasible for indoor environments.



(a) Radius Error (Closed-Form) **(b)** Radius Error (Closed-Form) **(c)** Radius Error (Corrected) **(d)** Radius Error (Labeled)

Figure 2: Radius correction and classification pipeline. (a) Histogram of errors using the closed-form log-normal shadow model to convert RSS to radius. (b) Boxplot of errors with respect to ground truth distance. Boxes represent the center 50% of the data. Red lines indicate the median value. (c) Radius error with respect to distance after applying our power regression to correct for skew. (d) Boxplot of non-noisy radius estimations (median denoted by a blue circle) overlayed with noisy predictions (denoted as a red +). Because our model identifies erroneous radius estimates, the average error is further reduced.

In [32], Redzic *et al*. use a wearable camera and wifi signals to localize a mobile person. They adopt a fingerprint-based approach consisting of: (i) images of the environment taken from the viewpoint of the person and (ii) RSS fingerprints at known calibration points. While the authors assume knowledge of the a priori distribution of the user's location, we do not make this assumption. In [33], the authors take this idea further and use SIFT features to assist with image-based localization. While both of these methods study localization with respect to the user and a mobile camera, our work focuses on localization with respect to a fixed camera. We propose to use the **W** modality to infer depth and combine it with RGB images to localize and track individuals.

3. From W to Ring Images

We aim to augment RGB data with **W** data to better locate and track individuals. To achieve this goal, we must formulate a relevant representation of **W** data to efficiently fuse it with RGB data. We propose an embedding which captures the radius estimation, error bounds, and confidence level (noise detection) for each antenna. We use a classification framework to infer the quality of the **W** data. This culminates in our embedded **W** representation, illustrated in Figure 3, which we call a *ring image*.

The proposed classification framework to infer the ring image is practical thanks to our RGB-W setup. We can automatically collect labeled data when a single person walks around the scene. As a result, we adjust the learned model automatically at test time to best fit the scene interferences.

3.1. Radius Estimation

For any individual i having **W** enabled, we observe the following information at a given time frame t in a space:

$$W_i^{(t)} = \{\mathrm{RSS}_1, ..., \mathrm{RSS}_j, \text{Phone mac}\}^{(t)}, \qquad (1)$$

where RSS_j is the received signal strength from antenna j. In a noise-free environment, RSS directly provides the radius (i.e., distance to the antenna) through a closed-form logarithmic expression such as the one presented by Chitte *et al*. in [9]:

$$\mathrm{RSS} = \mathrm{RSS}_0 - 10\beta \log_{10}\left(\frac{r}{r_0}\right) \qquad (2)$$

$$r = r_0 10^{(\mathrm{RSS}_0 - \mathrm{RSS})/(10\beta)}, \qquad (3)$$

where RSS_0 and r_0 are calibrated at a known reference point and β is the path loss exponent, typically real valued between 2 and 4 with larger values indicating more noisy environments. Our reference point is $r_0 = 1$ meter from the antenna and we use $\beta = 3.5$.

In reality, RSS is noisy and anisotropic, therefore Equation 3 is no longer suitable. Figures 2a and 2b show the error between the ground truth radius and the closed-form radius (r) from Equation 3. The closed-form radius has an average error of -1.2 meters since it does not model the environment interferences. It is clear that the log-normal shadowing model exhibits systemic skew. We propose a power regression to learn and correct the original radii, on a per antenna basis: $\hat{r}_j = (r_j/a_j)^{1/b_j}$ where r_j denotes the closed-form radius (Equation 3), $\hat{r}_j$ the corrected radius (see Figure 3), and a_j, b_j the fitted coefficients for antenna j. After applying our correction, the corrected average error (shown in Figure 2c) is -0.6 meters. Since the number of outliers is large, we propose a classifier to detect them.

3.2. Noise Detection

We suggest detecting noisy RSS readings to avoid introducing errors in our localization methods. We claim that having fewer "clean" RSS readings is better than more, but noisy measurements.

We apply a systematic method to infer the quality of the corrected radius to antenna j by modeling the joint responses of all antennas across a temporal window. All other antennas are used as a measure of coherence to validate the observation reported by antenna j. Our intuition is that in

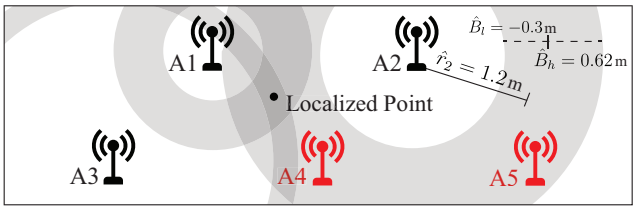


Figure 3: Illustration of the proposed ring image. Top view of the ground plane. Antennas A4 and A5 are classified as noisy (in red). The localized point is the weighted center of mass using the intersection of all non-noisy radii to antennas A1, A2, and A3 (in black). All distances denote meters.

the presence of noise, the estimated distances across the antennas are not coherent in space and time. Our classifier learns the subset of points that are coherent.

To train a classification model for each antenna, we compose a single feature vector $x^{(t)}$. It is important to note that $x^{(t)}$ is constant for each classification model but the label $y_j^{(t)} \in \{0, 1\}$ for antenna j at time t, varies. When $\hat{r}_j^{(t)}$ is more than 1.5 meters from the ground truth, we consider this noise and assign $y_j^{(t)} = 1$. Let $\hat{r}_j^{(t)}$ denote the corrected radius for antenna j at a given time t. Formally:

$$x^{(t)} = \{\hat{r}_1^{(t)}, ..., \hat{r}_\alpha^{(t)}, \hat{r}_1^{(t-1)}, ..., \hat{r}_\alpha^{(t-1)}, ...\}, \quad (4)$$

where α is the number of antennas. We train a support vector machine on $(x^{(t)}, y_j^{(t)})$ examples. The output of our classifier reduces the average radius error to 0.2 meters, as shown in Figure 2d. More detailed analyses are presented in Section 6.

3.3. Error Bounds Inference

Our goal is to convert **W** data into an image representing as much information as possible. In addition to the estimated radius, which can be represented by a circle on an image ground, we want to also model the expected error in our representation. In Figure 2d, we can see that the error bounds change with respect to the distance. Let $\hat{B}_l(\hat{r})$ and $\hat{B}_h(\hat{r})$ be respectively the estimated lower and upper bound on the estimated radius $\hat{r}$. This gives us the range:

$$\hat{B}_l(\hat{r}) < \hat{r} < \hat{B}_h(\hat{r}) \quad (5)$$

We propose to represent such range as a ring instead of a circle. Thanks to our training data (automatically collected with RGB-W; see Section 6), we learn a regression model to infer the radius error for each antenna. We use a support vector regression to estimate the error bounds ($\hat{B}_l$ and $\hat{B}_h$). This gives a "width" to each circle (see Figure 3). In the next section, we present the framework to jointly reason with RGB and ring images to locate and track individuals.

4. RGB-W Human Localization

We want to jointly use RGB with **W** data in a unified representation to locate individuals in the space. Intuitively, we believe that RGB can accurately estimate angular coordinates with respect to the camera center, whereas the **W** can provide an estimate of individuals' distance to the camera (depth), and better address ambiguities in the presence of occlusion.

We have intentionally represented the **W** data as ring images to leverage a sparsity driven formulation to locate individuals on the ground. In this section, we show how to naturally fuse foreground images from a camera and the ring images to infer the ground plane occupancy of individuals in the scene. We formulate the task as an inverse problem using a multi-modal dictionary and a cascade of convex solvers.

4.1. Problem Formulation

We aim to infer the location of individuals on the ground given foreground silhouettes from a single camera as well as incomplete RSS data, *i.e.*, RSS measured from a sub-set of individuals only. Both signals are noisy as illustrated in Figure 6. We frame this as a best subset selection problem:

$$\arg\min_x \|x\|_0 \text{ s.t. } Ax + n = b, \quad (6)$$

where x represents the discretized ground plane points, b the observed data (*i.e.* foreground silhouettes + ring images) at a given time, A a dictionary representing for each ground plane point the ideal expected observation, and n is the noise level. We want to find a sparse occupancy vector x that can reconstruct the observation b.

The key difference with previous work [17] is the building of a new dictionary A and observation b. We also propose a cascade of solvers to best leverage RGB-W data.

4.2. A Multi-Modal Dictionary

We want to represent the possible set of "ideal" observation of an individual occupying a ground plane point. We construct a dictionary, denoted as A, where each column, namely atom, represents the expected foreground image and the expected ring image. Dictionary A is of size $n \times m$, where n is the size of an atom (sum of foreground and ring image size) and m is the number of ground plane points (same dimension as x).

The foreground images are approximated with a binary rectangular shape. The ring images are made by summing the antenna responses. Each response from an antenna (a single ring) is a binary image. The final ring image is the pixel-wise sum of all the binary ring shapes. Since humans are approximated with rectangles or rings (*i.e.* no fine-grained information), both the foreground and ring images can be downscaled to 160x120 without any loss of accuracy. The proposed dictionary has the following properties:

Atom Linearity. In the presence of occlusion, the linear operation Ax in Equation 6 is wrongly summing the binary

foreground images to match the observed data. However, the ring images are correctly modeled as a linear operation. They indeed sum up to match the observed data. As a result, the multi-modal nature of the atoms can better handle occluded individuals.

Incomplete Atoms. The **W** data is sparse, *i.e.*, a subset of individuals might broadcast their data, and only a subset of the antennas might be used (classified as non-noisy). As a result, several ring images are possible given the occupancy of an individual on the ground. For each ground plane point, several columns are created in the dictionary as illustrated in Figure 5 (the last column $(i+3)$ is only made of the foreground silhouette to locate individuals who do not broadcast their phone signals).

4.3. Representing the Observation Vector

The observation vector b is the output of a background subtraction algorithm generating the binary image vector of foreground silhouettes augmented with the estimated ring image:

$$b = [-F - W]^T, \tag{7}$$

where F is the binary foreground silhouettes image, and W is the ring image. For instance, any column from dictionary A can be considered as an observation of a single individual (see Figure 5).

4.4. Cascade of Lasso and BPDN Solvers

Ideally, we want to solve Equation 6 which is a NP-hard problem. We propose to relax it by leveraging the multi-modal nature of our data.

Our multi-modal representation, and more precisely, the **W** modality, provides additional prior on the desired solution such as the lower bound of the number of individuals to locate. We propose to leverage that with a cascade of solvers. The occupancy vector x can be recovered by relaxing the formulation to a Basis Pursuit De-Noise problem:

$$x^* = \underset{x}{\text{argmin}} \frac{1}{2}\|b - Ax\|_2^2 + \lambda\|x\|_1, \tag{8}$$

where λ is the trade-off between sparsity level and reconstruction fidelity.

Several solvers exist for Equation 8 such as the Active Set Pursuit algorithm introduced by [15], a re-weighted scheme [1], or greedy approach [17] to efficiently approximate the solution. The quality of the solution is highly sensitive to the parameter λ. It actually depends on the estimated prior of the noise level and sparsity level.

Thanks to **W** data, and more precisely to the number of captured mac ID, we now have a minimum bound on the sparsity level. We propose to leverage that in the resolution of Equation 8 by solving a cascade of two solvers as described in Algorithm 1.

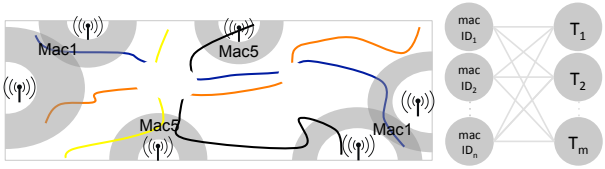


Figure 4: (Left-side) Top view of a collection of tracklets generated with high confidence and the collected W data (plotted as rings). (Right-side) Illustration of the bipartite graph to match mac ID_i to the tracklets T_i.

We first reformulate Equation 8 as a Lasso problem where the sparsity level is provided thanks to the number of observed mac ID (step 1 of Algorithm 1). Indeed, when the sparsity level is available, Lasso formulation is the natural formulation. The output of the Lasso solver might not locate people who have not sent a **W** data. Therefore, we solve the Basis Pursuit De-noising on the residual error to handle the missing detections using RGB only (step 3 of Algorithm 1). The Lasso formulation is looking for atoms that match the observed foreground and ring images. This reduces the number of candidate ground plane points.

Algorithm 1: Cascade of Convex ℓ_1 Solvers

Input: The dictionary A, observation signal b, ring image, and N the number of captured mac ID

Output: The occupancy vector x.

1. Solve Lasso formulation for RGB-W data:
   $x^{RGBW} = \underset{x}{\text{argmin}} \|b - Ax\|_2 \text{ s.t.} \|x\|_1 = N,$
2. Update b: $b = b - Ax^{RGBW}$
3. Solve BPDN for visual residual:
   $x^{RGB} = \underset{x}{\text{argmin}} \|x\|_1 \text{ s.t. } \|b - Ax\|_2 < \varepsilon,$
4. Final result: $x = x^{RGBW} + x^{RGB}$

In Section 6, we evaluate our cascade of Lasso and BPDN solvers against a single solver as well as previous work. In the next section, we show how to leverage W data to better track individuals across time.

5. RGB-W Human Tracking

Our eventual goal is to track humans in extreme conditions, *i.e.*, large crowded spaces given RGB-W data. Tremendous amount of works have addressed the multi-object tracking (MOT) problem, and more precisely, the well-known tracking-by-detection task [5, 1, 24, 10, 12, 21]. In brief, a directed graph is created where the nodes represent the detections across time and the edges encode the similarity cost. Global optimization algorithms exist to find

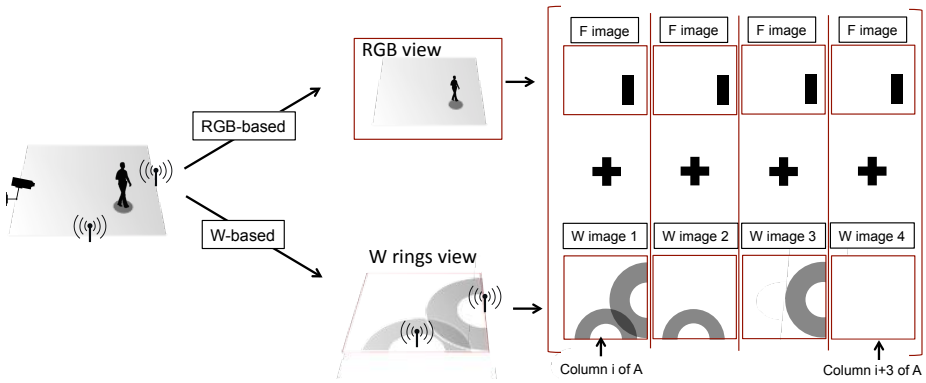


Figure 5: Illustration of dictionary construction given RGB-W data. For each ground plane point, the i^{th} column is made of the foreground ideal observation (F image) concatenated with the top view ring image of the **W** data for various antenna responses (**W** images i).

the best assignments with the Hungarian algorithm for online frame by frame mode, or k-shortest path /min cut max flow algorithm [24, 5] for batch mode. The real bottleneck remains in the similarity measure in specific "sensitive" cases, *e.g.,* when individuals interact and/or occlude each other. The "sensitive" cases can be detected by simply looking at the possible candidate targets. Reciprocally, we can connect detections that did not encounter "sensitive" cases, commonly referred to as tracklets (*i.e.*, short trajectories with high confidence). As a result, solving the tracking problem reduces to connecting these tracklets.

The nature of RGB-W data enables us to use a new source of information to reason on the similarity measure between tracklets. At irregular time frames, referred to as anchor points, we have access to a rough approximation about the locations of specific individuals thanks to the W data (see Figure 4). Therefore, we can assign a unique id (mac ID) to a subset of individuals to improve the data association algorithm by comparing the ids.

5.1. Assigning Mac ID to Tracklets

In order to improve tracking algorithm and offer the next generation of high precision location-based service, we aim to assign each mac ID to an observed tracklet. We formulate the data association problem as a bipartite graph.

Let $G_b = G_b(V_1 \cup V_2, E)$ where vertices V_1 represents the mac ID and V_2 represents the observed tracklets (see Figure 4). The weight d_{ij} of an edge $e_{ij} \in E$ represents the cost to assign the mac ID$_i$ to the tracklet j. We use Euclidean distance: $d_{i,j} = \|w_i - t_j\|_2$, where w_i is the center of mass[1] of the intersecting rings, and t_j the tracklet coordinate at the same time frame. We use the minimum weight bipartite matching algorithm presented in [20] to find the optimal assignment.

[1]For computing the center of mass, specific weights corresponding to the number of ring overlaps at a particular point can serve as additional model parameters.

6. Experiments

6.1. Data Collection

Our goal is to study the impact of complementing RGB streams with **W** data to locate and track individuals in crowded scenes. To the best of our knowledge, such RGB-W dataset does not exist. Therefore, we collected a new dataset of RGB-W data from both indoor and outdoor scenes where over ten individuals are simultaneously observed within the field of view of a single camera. At a density of 1 person/m^2, this leads to high levels of self occlusion (see Figure 6). The observed foreground silhouettes are noisy and highly ambiguous for occluded individuals. The **W** modality is measured with Beacon technology using one to four Beacons (antennas). Each person is equipped with an iPhone or Android device broadcasting the RSS to a server. To help promote additional research studies and additional data collection campaigns, the dataset, tools, hardware details, and code (including iPhone, Android, and server applications) are available online.[2]

Data collection, for both indoor and outdoor settings, is performed in challenging real world environments such as electronically-dense university buildings and outdoor courtyards.

For each frame and for each individual, the dataset includes annotated ground truth coordinates (in the xy ground plane), RSS to all antennas, closed-form radii values, and mac ID.[3] All cameras, servers, and phones are synchronized using a calibrated UNIX timestamp at the millisecond level. In the next section, we present the results of our RGB-W based algorithm to locate and track individuals.

6.2. Localization Results

We first study the performance of our RGB-W based localization method (Algorithm 1). We use the precision and

[2]http://vision.stanford.edu/rgbw/

[3]For iPhones, the advertising identifier (IDFA) is used.

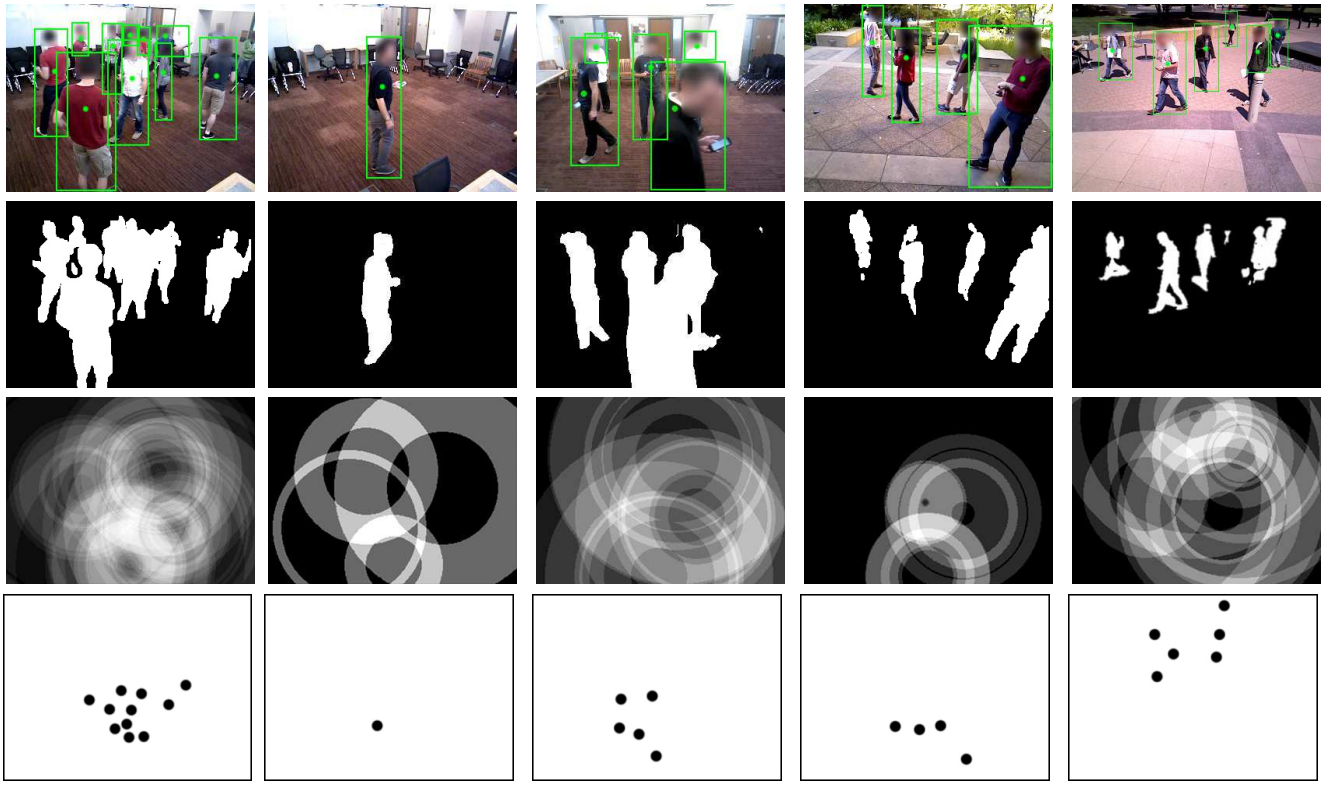


Figure 6: First row: original RGB image. Second row: extracted foreground silhouettes. Third row: superimposed ring images from all individuals. Whiter areas indicate regions which are likely to contain individuals. Fourth row: resulting RGB-W localization (top view). This does not necessarily correspond to the whitest ring image regions since we perform the optimization jointly with foreground silhouettes.

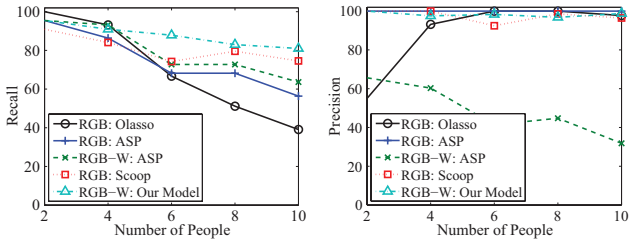


Figure 7: Precision and recall curves for several algorithms including our proposed RGB-W method.

recall as our primary performance metrics. A true positive is detection on the ground that is less than 1m away from the ground truth. We intentionally set it high enough to promote **W** based method. We compare our approach against the following baselines in Table 1.

RGB Only. The sparsity driven formulation was initially introduced for RGB camera to outperform previous work. Several solvers exist such as OLasso [1], Scoop [17], or ASP [15]. We compare our method against these methods using the same formulation but with a dictionary made of the foreground images only (without the ring images). The O-Lasso method performs best among the RGB only method but significantly less than our proposed RGB-W based method. Note that we also studied the impact of providing the number of observed mac ID as a lower bound to the number of individuals to locate. It did not increase their performance.

W Only. We evaluate the performance of **W** based method such as trilateration [27] and fingerprinting to get more insight on the localization error of the **W** data (without using our proposed ring images). Both methods perform poorly. We also evaluate our proposed ring-based representation to locate individuals (without using RGB data). It significantly outperforms the trilateration and fingerprint approaches.

Our Ring Model Only. Figure 8 illustrates the impact of our proposed method to generate the ring images. We can see that each step (described in Section 3), has a positive impact on the final localization error. Our full pipeline with the noise detection given temporal features reduces the average localization error to 0.81.

RGB-W. The results from Table 1 demonstrates that our method outperforms previous work, even with noisy **W** data. We study the impact of using our cascade of ℓ_1 solvers against a single solver such as the one proposed by

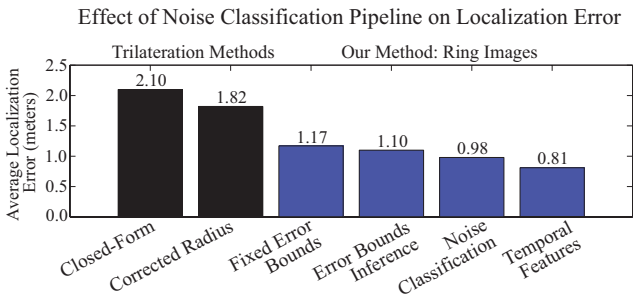


Figure 8: Localization accuracy at each step of our noise classification pipeline. Closed form and corrected radius errors were computed with trilateration. All other categories are computed with ring images.

[15] (referred to as ASP). Without using the cascade approach, our multi-modal dictionary with a known solver does not exhibit a gain in performance (the precision even degrades). Figure 6 illustrates some qualitative examples of a scene observed with RGB-W data. We also provide the recall/precision rate with respect to the number of individuals in Figure 7. As expected, the performance decreases with the number of individuals and the ranking of the methods stay the same.

Table 1: Performance evaluation of the localization task in terms of recall and precision with respect to other methods.

		Recall	Precision
RGB Only	O-Lasso [1]	45.6%	60.1%
	Scoop [17]	43.4%	43.2%
	ASP [15]	43.4%	39.2%
W Only	Trilateration [27]	27.2%	6.1%
	Fingerprinting	27.0%	9.2%
	Our model (ring-only)	40.5%	61.7%
RGB-W	ASP [15]	43.4%	26.6%
	Our model	**69.5%**	**72.7%**

Table 2: Impact of turning on the **W** modality. The first column represents the percentage of individuals who have the **W** signal on.

# W-Enabled Devices	Recall	Precision
100%	**82.0%**	**98.1%**
80%	78.0%	98.0%
60%	77.0%	99.0%
40%	75.0%	98.6%
20%	71.0%	94.5%
0%	74.0%	96.0%

In Table 2, we illustrate the impact of having the **W** modality available with respect to the number of people present in the scene. When a small subset of people (less than 20%) were broadcasting their **W**, performance slowly decreased. As soon as more than half of the subset of individuals are broadcasting their RSS, the average localization performance of the full system is improved outperforming RGB only approaches and **W** based approaches by significant margin.

Table 3: Performance of assigning the correct mac ID to an individual. The number of people indicates the number of people in a scene assuming all people are broadcasting **W** data.

Number of People	Greedy	Our Model
2	61.7%	64.0%
4	52.0%	57.2%
6	45.6%	53.4%
8	36.1%	45.3%
10	27.3%	30.2%
12	21.0%	28.6%

6.3. Performance of Assigning Mac ID to Tracklets

The RGB-W data enables the use of a new similarity measure to solve the tracking problem. We study the performance of assigning the mac IDs to detected individuals given their rough localization. Table 3 presents the performance of the assignment as a function of the number of individuals in the scene. During the experiments, individuals were moving in highly dense manner, *i.e.*, 1 to 2 meters away from each other even when two individuals were present. The assignment is based on minimizing the global distances between the detections from RGB-W and **W** only. In Table 3, we can see that our proposed method is outperforming the greedy approach but is still challenging. The success rate is not high. Future work can investigate on how to increase the performance of such task by comparing the temporal dynamics of the **W** with respect to the tracklets.

7. Conclusion

In this work, we suggested to fuse the vision and wireless modalities to solve a common problem, namely, human localization and tracking. In the past years, we have witnessed widespread deployment of affordable sensing devices to capture both visual and wireless signals. We have shown in this paper how to leverage these multi-modal sources of data into a unified framework. We demonstrated that it is possible to improve the localization and tracking of individuals in dense, crowded scenes with a single monocular camera by complementing it with wireless data.

Acknowledgements. We would like to thank Roland Angst, Alexandre Robicquet, Andre Esteva, and the entire Stanford Vision Lab for assistance with data collection. Alexandre Alahi is funded by the Swiss National Science Foundation under fellowship number P300P2_154535.

References

[1] A. Alahi, L. Jacques, Y. Boursier, and P. Vandergheynst. Sparsity driven people localization with a heterogeneous network of cameras. *Journal of Mathematical Imaging and Vision*, 2011.

[2] A. Alahi, V. Ramanathan, and L. Fei-Fei. Socially-aware large-scale crowd forecasting. In *CVPR*, 2014.

[3] P. Bahl and V. N. Padmanabhan. Radar: An in-building rf-based user location and tracking system. In *Joint Conference of the Computer and Communications Societies*. IEEE, 2000.

[4] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool. Pedestrian detection at 100 frames per second. In *CVPR*. IEEE, 2012.

[5] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua. Multiple object tracking using k-shortest paths optimization. *PAMI*, 2011.

[6] J. Biswas and M. Veloso. Wifi localization and navigation for autonomous indoor mobile robots. In *Robotics and Automation, IEEE Intl. Conference on*. IEEE, 2010.

[7] J. Biswas and M. Veloso. Depth camera based indoor mobile robot localization and navigation. In *Robotics and Automation, IEEE Intl. Conference on*. IEEE, 2012.

[8] Y.-C. Cheng, Y. Chawathe, A. LaMarca, and J. Krumm. Accuracy characterization for metropolitan-scale wi-fi localization. In *Mobile Systems, Applications, and Services, Intl. Conference on*. ACM, 2005.

[9] S. D. Chitte and S. Dasgupta. Distance estimation from received signal strength under log-normal shadowing. In *Signal Processing, Intl. Conference on*. IEEE, 2008.

[10] D. Delannay, N. Danhier, and C. D. Vleeschouwer. Detection and recognition of sports(wo)man from multiple views. In *Distributed Smart Cameras, Conference on*. IEEE, 2009.

[11] M. Enzweiler and D. M. Gavrila. Monocular pedestrian detection: Survey and experiments. *PAMI*, 2009.

[12] R. Eshel and Y. Moses. Homography based multiple camera detection and tracking of people in a dense crowd. In *CVPR*. IEEE, 2008.

[13] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Cascade object detection with deformable part models. In *CVPR*. IEEE, 2010.

[14] B. Ferris, D. Fox, and N. D. Lawrence. Wifi-slam using gaussian process latent variable models. In *Intl. Joint Conferences on Artificial Intelligence*, 2007.

[15] M. P. Friedlander and M. Saunders. Active-set methods for basisi pursuit denoising. Sep 2008.

[16] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*. IEEE, 2014.

[17] M. Golbabaee, A. Alahi, and P. Vandergheynst. Scoop: A real-time sparsity driven people localization algorithm. *Journal of Mathematical Imaging and Vision*, 2014.

[18] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox. Rgb-d mapping: Using depth cameras for dense 3d modeling of indoor environments. In *Intl. Symposium on Experimental Robotics*. Citeseer, 2010.

[19] J. Hightower and G. Borriello. Location systems for ubiquitous computing. *Computer*, 2001.

[20] J. Hopcroft and R. Karp. An nˆ5/2 algorithm for maximum matchings in bipartite graphs. *SIAM Journal on Computing*, 1973.

[21] S. M. Khan and M. Shah. Tracking multiple occluding people by localizing on multiple scene planes. *PAMI*, 2009.

[22] Z. Khan, T. Balch, and F. Dellaert. Mcmc-based particle filtering for tracking a variable number of interacting targets. *PAMI*, 2005.

[23] J. Krumm and E. Horvitz. Locadio: Inferring motion and location from wi-fi signal strengths. In *Mobile and Ubiquitous Systems, Intl Conference on*, 2004.

[24] L. Leal-Taixe, G. Pons-Moll, and B. Rosenhahn. Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. In *ICCV Workshops*. IEEE, 2011.

[25] J. Letchner, D. Fox, and A. LaMarca. Large-scale localization from wireless signal strength. In *National Conference on Artificial Intelligence*. AAAI Press, MIT Press, 2005.

[26] C. Liu, T. Scott, K. Wu, and D. Hoffman. Range-free sensor localisation with ring overlapping based on comparison of received signal strength indicator. *Intl. Journal of Sensor Networks*, 2007.

[27] H. Liu, Y. Gan, J. Yang, S. Sidhom, Y. Wang, Y. Chen, and F. Ye. Push the limit of wifi based localization for smartphones. In *Mobile Computing and Networking, Intl. Conference on*. ACM, 2012.

[28] T. Miyaki, T. Yamasaki, and K. Aizawa. Tracking persons using particle filter fusing visual and wi-fi localizations for widely distributed camera. In *Image Processing, IEEE Intl. Conference on*. IEEE, 2007.

[29] V. Moghtadaiee and A. G. Dempster. Wifi fingerprinting signal strength error modeling for short distances. In *Indoor Positioning and Indoor Navigation, Intl. Conference on*, 2012.

[30] E. Mok and G. Retscher. Location determination using wifi fingerprinting versus wifi trilateration. *Journal of Location Based Services*, 2007.

[31] P. Nillius, J. Sullivan, and S. Carlsson. Multi-target tracking-linking identities using bayesian network inference. In *CVPR*. IEEE, 2006.

[32] M. Redzic, C. Brennan, and N. E. O'Connor. Dual-sensor fusion for indoor user localisation. In *Multimedia, Intl. Conference on*. ACM, 2011.

[33] A. J. Ruiz-Ruiz, O. Canovas, R. A. R. Munoz, and P. E. L.-d.-T. Alcolea. Using sift and wifi signals to provide location-based services for smartphones. In *Mobile and Ubiquitous Systems*. Springer, 2012.

[34] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore. Real-time human pose recognition in parts from single depth images. *Communications of the ACM*, 2013.

[35] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on riemannian manifolds. IEEE, 2008.

[36] V. Vivekanandan and V. W. Wong. Concentric anchor beacon localization algorithm for wireless sensor networks. *Vehicular Technology, IEEE Transactions on*, 2007.

[37] P. A. Zandbergen. Accuracy of iphone locations: A comparison of assisted gps, wifi and cellular positioning. *Transactions in GIS*, 2009.