# Using Relaxation to Fuse RFID and Vision for Object Tracking Outdoors

Rana H. Raza[*,†] and George C. Stockman[*]
Michigan State University, East Lansing, MI 48824 USA

## Abstract

Fusion of Radio Frequency Identification (RFID) with Computer Vision (CV) can significantly improve performance in applications of autonomous vision and navigation, activity analysis, site monitoring, and especially in outdoor environments. RFID and CV provide both overlapping and unique information for deciding on object identity, location, and motion. We use relaxation to control the integration of information from CV, RFID, and naïve physics. Work site analysis must proceed even when information from one sensor or information source is unavailable at some time instances. Simulations show how fusion can greatly increase tracking performance while also reducing computational cost. Test cases show how fusion can solve some difficult tracking problems outdoors.

**Key Words**: Tracking, stereo, RFID, fusion, relaxation, site monitoring.

## 1 Introduction

We are interested in application problems in site monitoring, security, and activity analysis. Examples are tracking both baggage and people in airports; workers, materials, and machines in construction sites; or patients and care workers in medical care facilities. There are many other important applications. We believe we are the first to report experimental work on fusion of RFID and CV for object identification and tracking in an outdoor environment.

Figure 1 shows our outdoor test site: we are studying how well we can detect and track RFID tagged moving objects. The site is a courtyard roughly *40 Sq m* with several stone posts, sidewalks, and several large trees that limit both movement and observation via sensors. The surrounding building has many corners and windows. We surveyed landmarks using a combination of tape measure, laser range finder, and then stereo vision once enough calibration features were available. The same landmarks were available for calibration of a set of RFID readers as well.

### 1.1 Basic Functionality Required

The applications of interest require a system that provides some or all of the following basic functions.

a. **Detection** of the presence of objects of interest (persons, machines, materials, vehicles…).
b. **Identification** of the objects (by class or by a unique object instance).
c. Object **location** in workspace coordinates or by designated areas.
d. Object **track**, if the object is moving.
e. Important **object properties**, such as shape, color, weight, speed, ownership, supplier, etc.
f. A memory **representation** of space and time including location of objects, trajectories, and behaviors.
g. Application specific processes that manage objects and control their behavior (such as collision avoidance or creation).

Fusion of CV and RFID may provide the base functionality. Higher level problem-specific analysis applied on top of functionality (*a*) to (*e*) can create a dynamic inventory of a workspace, infer what agents or objects are doing (*function f*), actively manage interactions, define and summarize events, etc. (*function g*).

### 1.2 Some Advantages and Disadvantages of CV

Computer Vision has been very successful in controlled indoor environments, but challenged in uncontrolled outdoor environments. The CV literature contains thousands of reports on object detection and recognition, tracking, and motion analysis. Image sensors are passive, cheap, can be far-seeing, and can collect a good deal of information about a scene. Commodity cameras easily produce frame rates useable for most human motion analysis. Detections and relationships in a 2D image can often be mapped to the real 3D scene. Using multiple camera stereo, objects can be located in 3D(or, special active sensors can yield range/depth images).

Object identification via CV -- function (*b*) above -- is often difficult and is usually based on sensed features (*e*). Even accurate features may not precisely identify an object, e.g., who is that person or what year is that Chevy Cruz

[*] College of Engineering, Engineering Building, 428 S. Shaw Lane.
[†] E-mail: razarana@msu.edu; Phone 1 517 325-3260; ranahammadraza.com.

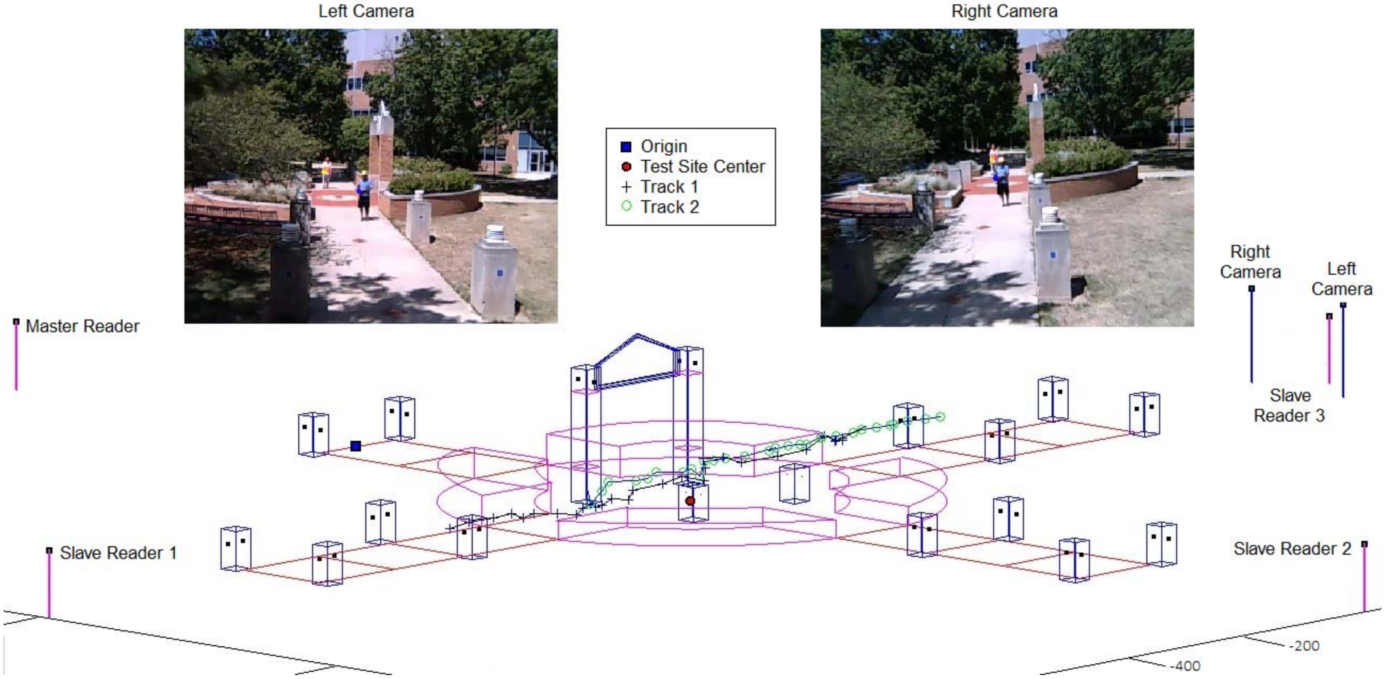


Figure 1: Outdoor workspace: about *40 Sq m* with trees and structures and a surrounding building

Identifying an object by features, if possible and reliable, can be computationally costly given the necessary signal processing and the variation of appearance over many possible 3D poses. So, while person identification using carefully imaged biometrics might yield identity accuracies of over 98 percent, more general object recognition via color camera image is far less accurate. Acquiring quality images under occlusion and variations in lighting also cause serious problems in CV applications. Uncontrolled outdoor environments might be dark, dusty, have rain or snow, and have both static and dynamic objects occluding the sensor view. Finally, CV may sometimes give too much information; for example, humans do not want to be imaged in private spaces, and may resent being watched in work places.

### 1.3 Some Advantages and Disadvantages of RFID

RFID can easily and reliably provide a unique object identification by transmitting a digital signal to a reader. With enough power, RFID can also transmit non visual features of an object, such as weight or ownership. RFID can operate in smoke, and darkness, thus it is widely used in sales and inventory systems and is replacing bar coding when cost permits. Object identification approaches 100 percent accuracy in commercial applications where objects are close to (presented to) a reader in a controlled environment. Objects with RFID tags can actually transmit their own physical description to an automated system or a security person. In robotics or material handling, the description might send a CAD model to a CV system to teach it how to recognize the object. RFID technology offers a wide variation in terms of cost, size, sensing distance, memory and processing power, and security [6]. With higher RFID frequency, higher data rate can be achieved. The higher data rate with appropriate anti-collision algorithm can enable a single reader to read a large population of tags. RFID can even be used to locate objects. In one method, a networked grid of short range RFID readers can be polled to find out which reader is sensing the nearby object. Alternatively, multiple long range readers can locate an object with an active tag by triangulation or trilateration [22]. The Real Time Location System [3] that we have used for our RFID sensing is discussed in Section 3. RFID codes cannot be sensed by humans and hence can yield less overt identity and might be tolerated in private human spaces.

RFID requires that an object be physically tagged, thus changing the object itself and requiring that the object be "cooperative". Although passive RFID tags can be tiny and do not require their own energy source, they are used for limited range and have limited memory. Highly functional RFID tags require an energy source for communication, memory, and processing. An RFID tag is a proxy for an actual object so it or its communication can be counterfeit, thus making an object appear to be what it is not. Simple examples of this would be one driver stealing an EZ-PASS device from another driver, a shopper moving a tag from a cheap article to an expensive one, or two airline passengers swapping their RFID boarding passes. Thus RFID tags in critical security applications need to use encryption and secure operating system principles. Fusion with CV can enhance security.

### 1.4 Outline of the Paper

Section 2 describes a few related prior applications and methods using fusion of RFID and CV. Section 3 describes the model of the problem we approach and Section 4 highlights the relaxation scheme as its solution. Section 5 gives critical test scenarios illustrating the benefits of fusion to object tracking and further analysis. Section 6 is our concluding discussion.

## 2 Background

We have surveyed many different projects dealing with fused CV and RFID. All of the projects using fusion were applied in indoor controlled environments. Here we describe just a few significant projects that used RFID and CV and demonstrate the potential for many other applications.

### 2.1 Monitoring People in a Controlled Indoor Environment

Fusion of RFID and CV has been used in a day-care environment by S. Nakagawa, et al. [13]. Parents can view their child's activity via the Internet. RFID tags are placed on play objects and on children so that readers in the play space can identify and locate them. The system can then select appropriate cameras for good views of selected children and/or objects. Software alarms can be implemented for interaction between special pairs of objects and summarization of an entire day's activity of a child can be done. The parents can know who or what their child interacted with in a given day and can locate video segments of these interactions. There are other applications requiring similar functionality – elder care, studying how shoppers examine items for sale, or how visitors examine art in a museum. Teixeira, et al. [21] reported tracking patients in indoor elder care using cameras on the ceiling and cell phones. Statistical methods were used to correlate the image features with the locations reported by the cell phones system.

### 2.2 Robotics Applications

Passive stereo vision can locate detected objects in a 3D volume [19]. An RFID reader can be used to identify an object observed in some 2D image, thus aiding stereo; or, a network of RFID readers can provide coarse 3D location without cameras. RFID technology also enables smart objects to communicate information about themselves not available to optical sensors; for example, object weight, container content, etc. A tagged rigid object can even help provide an optical observer with a network downloaded CAD model of itself to be used for pose computation by the observer. This was done by Hontani, et al [7], who also used visual tags on objects as a starting point in matching an observed image of the object to a projection of the 3D CAD model. Chae, et al [2] using fusion of RFID and CV, proposed a global to fine localization algorithm for a mobile robot in an indoor environment. The problem of dynamic obstacle recognition in mobile autonomous platforms is addressed in [10] by using fused information. Limitations in mobile robots such as dead reckoning or wheel slippage are also addressable using fusion. Moreover, information can be written to an active tag perhaps recording object location and time, or perhaps what operation or measurement was done on the object.

### 2.3 Tracking Elephants in the Dallas Zoo

The commercially available Real Time Location System (RTLS) [3] can both identify and locate an active tag in a work or play space indoors or outdoors. Typically, five or more readers are distributed about the space, which can be as large as *600 Sq m*. An implementation at the Dallas Zoo provides displays to visitors that locate each elephant in a 2D map of their area [5]. Elephants wear an active tag on their ankle. The system can locate the tag (elephant) within *1* or *2 m* accuracy and update the location every *2 sec*. This is not an application of fusion with CV; however, cameras could easily be added and used as in the aforementioned day care application. It would be simple to place cameras about the elephant area and then add them to the area map so that they can be used to observe a particular elephant with a known location in the map. We use RTLS in our outdoor research on fusion and this RFID data is illustrated in Section 4. Evaluations of the used RTLS in both indoor and outdoor environments are available on the internet [9]. On a similar note an airport security system has been proposed and analyzed by Zekavat, et al. [24]. It is primarily based on RFID for location and tracking of passengers and staff.

### 2.4 E-Passports

Electronic passports have been designed that combine RFID tags with conventional printed information and a photograph [8]. The RFID tag can privately and quickly transfer information about the person (object) to a machine, thus streamlining information transfer and saving personnel time. Encrypting techniques can make forgery much more difficult. There must be a digital photo or fingerprint on the tag, which can be compared to the live person. Once a person has entered a secure area, the digital photo or fingerprint can be compared to live biometrics taken in the workspace, such as in doorways or stations, to verify the location or activity of the person. Early on, some reengineering of the E-passport had to be done to shield the RFID tag so that it could not be read by hackers in unofficial places when the passport was just slightly open.

### 2.5 Sensor and Cell Phone Networks

Wireless sensor networks are networks of compact, cost effective nodes that sense and communicate environmental conditions such as light and temperature etc. Many applications use sensory tags that are RFID tags, which incorporate sensory functionality in addition to identification and possibly localization. Sensor networks alone can be used to provide location information using relative location between sensor nodes [11]. In applications where location monitoring is required, sensor nodes are oriented with respect to a global coordinate system so as to provide geographically meaningful data.

Functionally cell phones are similar to active RFID tags and cellular towers are similar to RFID readers operating over large distances. Commodity pricing brings impressive power to cell phones at moderate cost. The latest smart phones have many useful sensors onboard, such as cameras, accelerometers, gyros, compass, GPS receivers, proximity and light sensors. Also they have large memories and exchange arbitrary data across networks. These sensors can be used in a local setting to compute position and movement information, for example a high-accuracy acoustic based ranging system using mobiles [15]. It is reported in [1] that NTTDoCoMo has manufactured cell phones with built in RFID modules. The present and future needs of these systems being fused together will generate a single efficient device useful for the applications cited in this paper.

### 2.6 Using Naïve Physics in Tracking

Many studies have focused on processing video to compute features used to extract and track moving objects. For instance, authors in [25] have used a vision based image registration algorithm to compensate for camera motion and then consecutive frames are transformed to the same coordinate system to solve the feature point correspondence problem. Zhou, et al. [26] have provided a blob tracking and classifying method in an outdoor environment by establishing correspondence between the object in view and the matched templates. A method of background subtraction and shadow detection in videos is provided in [4]. Meier, et al. [12] have reported an algorithm that can automatically extract moving objects from an image sequence. A survey of passive monocular methods is given by Veenman, et al. [23]. Consistency of color, texture, shape, and motion can be used to track an object region across multiple video frames. Variable lighting, variable 2D projections of a 3D object, and occlusion of one object by another present difficulties. Applications that need to recognize what the objects are face additional uncertainty and complexity. For example, an autonomous vehicle needs to identify obstacles in its path using their image extent and their motion or apparent motion [14]. Passive tracking using only cameras remains an important area of continued research.

## 3 Problem Representation

Our problem is to compute in real time the trajectories of *N* objects moving within a known 3D workspace. Diverse sensor observations are combined into object locations, and possible identities, at discrete time steps, which must be aggregated into *N* object trajectories. Without loss of generality, we may ground our discussion using the Site Safety System (SSS), where we want to track workers, materials, and machines in a work site. We abstract the information structures in order to support a system with diverse information sources and constraints and processes that may not have knowledge of each other.

### 3.1 Tokens Code Observations from Images and RFID

Sensor observations, and combinations of them, produce ***tokens*** $\boldsymbol{\tau}=<x, y, z, t, v, L>$, each coding that an object with identity, or name, *L* and feature vector *v* is at location *(x,y,z)* at time *t*. Some tokens will have incomplete information: for example, Identity *L* may be absent from camera observations and object features may be absent from RFID observations. 3D coordinates may be absent for an observation from a single camera image or single RFID reader. Two or more of these tokens can be combined in the processing to get 3D coordinates.

### 3.2 Object Tracks {<x, y, z, t, v, L>}

The site safety system SSS must detect, identify, and locate objects in a few video frames *k*. SSS may know $L = f(<x, y, z, v, t>)$ from sensor subsystems that use RFID or visual features. SSS can also use “tracking” to determine the label $L = f(\{<x, y, z, v, t\text{-}k>\})$ based on prior tokens or perhaps even forward tokens $\{<x, y, z, v, t+k>\}$. If object identity *L* is known, other object features $w = f(L)$ may be available from an RFID tag, such as object mass or CAD model. Finally, we note that if sensors supply object speed or acceleration, as cell phones may, we consider these as components of *v* along with color, texture, elongation, etc. of its image.

An **object track** is *k* or more tokens in time sequence with consistent object identity and features that also satisfy constraints for motion in space. Tracking is an important concern of this paper, and is a low level of motion understanding that uses naïve physics to aggregate observations over time. Heuristics from naïve physics enable aggregation of individual tokens into a sequence or track, one for each moving object. As objects move through the workspace, they may be occluded at any instant from either cameras or RFID readers so there may not be multiple tokens to fuse. Smoothness constraints, or motion applied over multiple time steps can be used to interpolate.

As we will see, it is not possible to assign unique object identities to every token at every time instance. Consider, for example, the popular shell game where a bean is placed under one of three shells that look alike [20]. When the shells are shuffled quickly in space, most people cannot track the shell containing the bean. If the shells are of distinct colors, then the problem of picking the final shell is easy. If the shells are identical in appearance, but the bean is an RFID tag, RFID readers are unlikely to be able to distinguish the tagged shell in space when the shells are close to each other. Consider three workers with hard hats each with a tag and close together; if the hats are the same color, RTLS cannot distinguish them; if we know which colors contain which tags and the hats are of different colors, the system can solve the matching problem and locate each hat within the CV distance error. In order to model ambiguity, we will have to allow multiple labels *L* in the tokens of an object track: these labels record the ambiguity of idenity at this point in time and space.

### 3.3 Obtaining 3D Object Location (x, y, z)

All sensors are calibrated to the same 3D workspace. One fundamental sensing concept is that a sensor observes an object along a ray in 3D space. The object can be located by intersecting two (or more) rays (or error cones) as done by using the standard stereo solution [19, Ch 13]. Possibly, two RFID readers, or one RFID reader and one camera can be combined this way as well. The underlying geometry is angle-side-angle, where the side is the known 3D baseline between the two sensors. A second fundamental sensing concept is where the sensor observes an object at some distance *d*. If the object transmission is observed by four such sensors, it can be located by trilateration, intersection of four spheres with radii equal to the sensed distances. An object can also be located by intersection of the ray/cone determined by an image observation and the spherical shell determined by distance *d* sensed by a single RFID reader. The commercial RTLS system encapsulates multiple RFID readers and yields a token with unique object identity and *(x, y)* coordinates on the ground plane of the workspace.

### 3.4 Fusion

We define fusion as the combination of different sensor tokens to obtain a token containing information from the different sensors or with new information computed from the tokens from the different sensors. Most importantly, RFID and CV tokens will be fused to combine object identity with object features and to provide or to refine object location.

### 3.5 Naïve Physics

Naïve, or common sense, physics provides many constraints about the world. These constraints have diverse forms and can be used in the filtering processes of relaxation. An object *n* must be at one and only one place at time *t* and location *<x,y,z>* can accommodate at most one object at time *t*. Object *n* is likely to have consistent form and visual features and observations of object *n* must be consistent with its identity. The motion of object *n* is likely to have smooth direction and velocity. Such constraints extend those used by Sethi and Jain [18] and Veenman, et al. [23]. Unfortunately, none are hard constraints! For instance, it may actually be that two objects have the same coordinates – when a driver enters a vehicle, for example.

### 3.6 Sensor Error and Synchronization Problems

In an outdoor environment we evaluated positional accuracy and reliability for the RFID based Real Time Location System (RTLS) and the stereo computation obtained using commodity cameras [17]. The calibration of stereo system and RTLS system were done on the same test site. An adequate number of distant points (in the background) and nearby points (in the foreground) were acquired to serve as calibration markers for stereo computation. The stereo infrastructure provided RMS positional accuracy of *19.3 cm* for *x,y* and *z* directions. The reported location accuracy for the RTLS system for static tags is *~1.5 m* and for dynamic tags is *2 m ~ 2.6 m*. RMS error does not include the occasional outliers that are possible from incorrect stereo correspondence or multiple path effects in RFID.

One significant practical problem for fusion is the different sampling rates of the sensors or the extended time needed to smooth data or to make decisions about the motion of an object. Due to time division multiplexing, our RFID system provides data on all objects every *2 sec*, while our stereo implementation could produce *10* updates per second for a few objects. In our experiments we typically force a common sampling time for RFID and CV and look back two time samples to estimate motion. The uncertainty of location for RFID is much larger than for CV for static objects and even larger for moving objects due to under-sampling. Interpolation using CV locations can be used with sparse RFID samples with reliable identity. Finally, it is possible that an object is invisible at some time steps to either or both CV and RFID due to occlusion.

## 4 Relaxation Labeling Scheme

We use discrete relaxation to create the tracks of the *N* objects and to update the time tokens comprising each track. Using relaxation, different sensors and sources of information can be turned on or off for experimentation or for practical reasons at a site. Fusion processes operate on a blackboard containing the set of tokens. When an observation is made, its initial label set is the set of all possible *N* known objects. Filtering processes are then applied to eliminate labels inconsistent with constraints. Sensing continues over the *T* time steps and naïve physics processes aggregate object consistent tracks.

For clarity, suppose that *N* objects are detected at time *t=1* and that we arbitrarily label these objects *1,…,N*. At time *t=2*, we have another *N* observation and we want to label each of those with the labels from time *t=1*. A label possible for a token at time *t=2* will be consistent in color, motion, and RFID identity with the tokens at time *t=1*. Initially, a new observation may detect any of the known objects, so all labels *L* are possible. A totally new object entering the site could be given a new unknown label. Most of these labels are filtered out quickly by failing constraints. For example, suppose *5* orange hard hats are detected at *t=1* and these have initial labels *3,4,6,8,9*. For any token for time *t=2* that is not orange, labels *3,4,6,8,9* will be deleted from its possible label set. Filtering can be done by space as well as by color. If any token at time *t=2* is unreasonably far from a token *m* at time *t=1*, then label *m* should be deleted from its label set.

### 4.1 Sensor Processes

A CV sensor process takes a video frame, segments it, and creates a token for tracking. Image features and two points on the imaging ray are stored in the feature vector *v*. Object *L* is initially unknown. An RFID reading produces a similar token, except that an object label *L* is known in almost all cases.

### 4.2 Combination Processes

Fusion processes take the sensor tokens and possibly merge information using ray intersection, ray-surface intersection, etc., whichever applies, and outputs a token with refined 3D location or label information. Filtering processes eliminate unlikely token labels by comparing tokens and by looking at feature vectors over time. (Our current software implementation of relaxation inputs combined tokens that have been precomputed from stereo correspondence. Similarly, RFID tokens have 3D information from the encapsulated RTLS system.)

### 4.3 Tracking Process

Naïve physics constraints are used to filter out highly unlikely labels for objects at time *t* based on the recent history of objects continuing from the *k* previous time steps. Our current results have used the current and *2* previous time steps.

### 4.4 Relaxation Labeling Algorithm

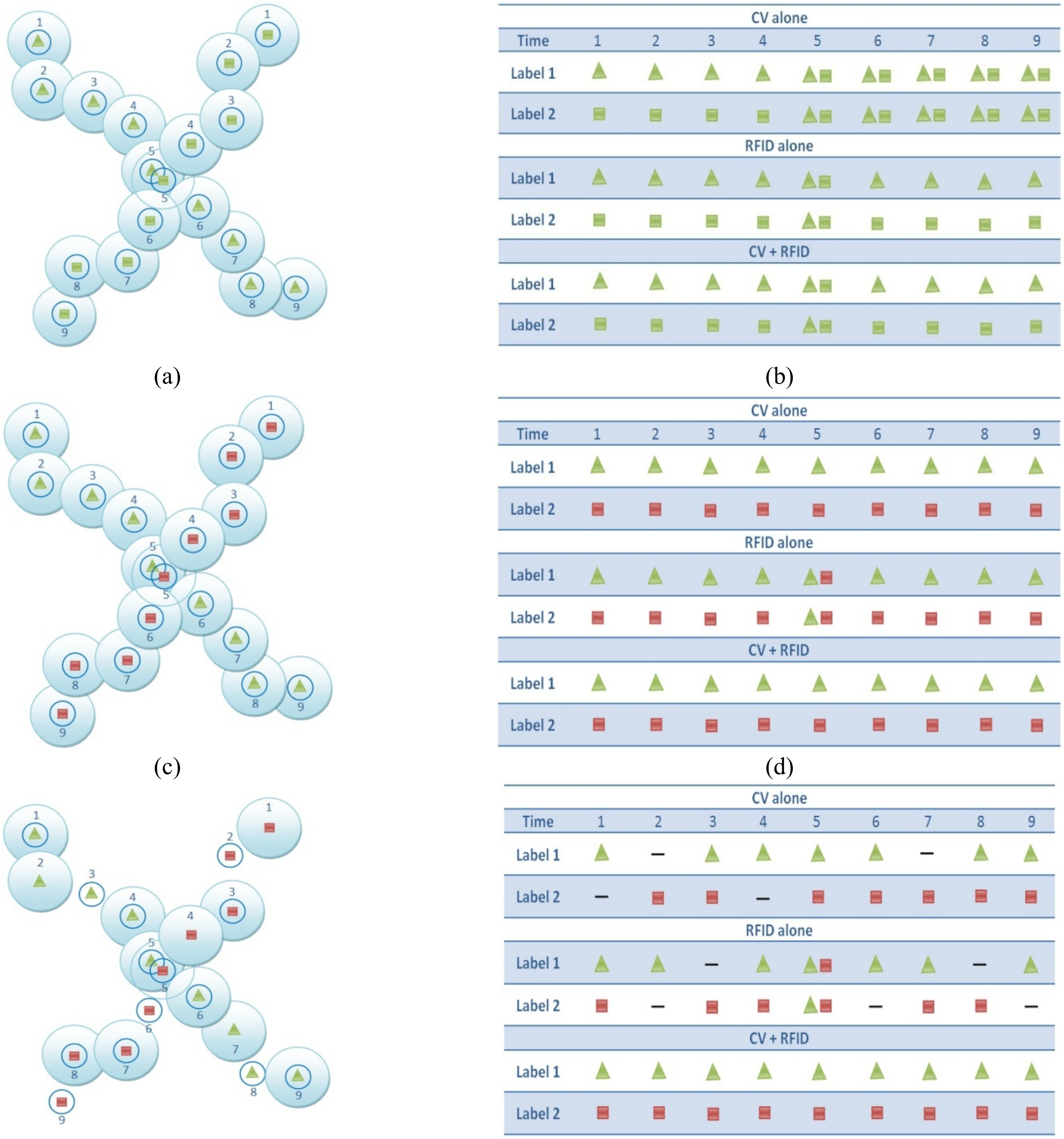
```
Output:  Object Labels $L_k$ and 3D location $XYZ_{Refined} \in R^3$
Input:   Object Labels $L_{k-2}$ and $L_{k-1}$ with color, RFID and $XYZ_{RFID}$ and $XYZ_{Stereo} \in R^3$

FOR      $t = k : K_{frames}$
    • Obtain color information if any for $XYZ_{Stereo}$ observations from 2D histogram matching
    • Sort colors into groups                         /* How many colored hats and balls and which colors*/

    Detect
    • $n$ number of $XYZ_{Stereo}$ observations detected     /* Motion detection and color detection*/
    • $m$ number of $XYZ_{RFID}$ observations detected       /* Active RFID*/
    • Generate empty label matrices for $p$ observations    /* $p = max(n,m)$*/
    • Assign $p$ labels to all $p$ observations and proceed to next pass

    Identify                                          /* Binary relationship criteria*/
    • Identify $XYZ_{Stereo}$ observations based on color information
    • Identify $XYZ_{RFID}$ observations based on identity
    • Correlate identity information

    IF        Only one color group                    /*All $XYZ_{Stereo}$ observations have same color*/
              No label elimination and proceed to next pass
    ELSEIF          Different color groups            /*Some $XYZ_{Stereo}$ observations have different color*/
              Eliminate labels from $p$ label matrices based on respective color groups and proceed to next pass
    END IF

    Locate                                            /* Binary relationship criteria*/
    • Set stereo and RFID location threshold values   /*Thresholds are defined based on sensor location accuracy and
                                                         object speed*/
    • Locate $XYZ_{Stereo}$ observations
    • Locate $XYZ_{RFID}$ observations with identity and location
    • Correlate location information

    Smooth
    • Calculate direction of flow/velocity for every $XYZ_{Stereo}$ observation at $t = k$ relative to $k-2$ and $k-1$
                                                      /*  z dimension gives valuable information here */
    • Correlate with RFID label/s identity and location information from $XYZ_{RFID}$

        IF            No difference in flow detected
                      Labels kept
        ELSEIF        Difference in flow detected     /* Only compatible labels remaining.*/
                      Eliminate unlikely labels
        END IF

    Compatible label/s obtained                       /*All possible labels for specific object*/
    • Compatible label/s provided to optimization process to obtain $XYZ_{Refined}$
END FOR
```

## 5 Test Cases and Analysis

We illustrate in this section how CV and RFID supplement each other in critical test cases. For clarity we first assume that for case I and II, both CV and RFID feeds are continuously available and the objects are not occluded by each other or the background and are moving with approximately the same velocities. Actual observations from our outdoor test site are used.

### 5.1 Test Cases to Explain Fusion

For case I, consider two objects represented as ▲and ■ with 3D data at each instance over *9* time frames. For better visualization the tracks are displayed in the *xy-plane* in Figure 2a. The objects are converging from north to south towards each other, and intersect at time frame *5* and thereafter follow their direction of motion without any change. Even if the objects were moving in a straight line, the points would appear



(a)

| CV alone | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Time | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Label 1 | ▲ | ▲ | ▲ | ▲ | ▲■ | ▲■ | ▲■ | ▲■ | ▲■ |
| Label 2 | ■ | ■ | ■ | ■ | ▲■ | ▲■ | ▲■ | ▲■ | ▲■ |
| RFID alone | | | | | | | | | |
| Label 1 | ▲ | ▲ | ▲ | ▲ | ▲■ | ▲ | ▲ | ▲ | ▲ |
| Label 2 | ■ | ■ | ■ | ■ | ▲■ | ■ | ■ | ■ | ■ |
| CV + RFID | | | | | | | | | |
| Label 1 | ▲ | ▲ | ▲ | ▲ | ▲■ | ▲ | ▲ | ▲ | ▲ |
| Label 2 | ■ | ■ | ■ | ■ | ▲■ | ■ | ■ | ■ | ■ |

(b)

(c)

| CV alone | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Time | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Label 1 | ▲ | ▲ | ▲ | ▲ | ▲ | ▲ | ▲ | ▲ | ▲ |
| Label 2 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |
| RFID alone | | | | | | | | | |
| Label 1 | ▲ | ▲ | ▲ | ▲ | ▲■ | ▲ | ▲ | ▲ | ▲ |
| Label 2 | ■ | ■ | ■ | ■ | ▲■ | ■ | ■ | ■ | ■ |
| CV + RFID | | | | | | | | | |
| Label 1 | ▲ | ▲ | ▲ | ▲ | ▲ | ▲ | ▲ | ▲ | ▲ |
| Label 2 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |

(d)

(e)

| CV alone | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Time | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
| Label 1 | ▲ | – | ▲ | ▲ | ▲ | ▲ | – | ▲ | ▲ |
| Label 2 | – | ■ | ■ | – | ■ | ■ | ■ | ■ | ■ |
| RFID alone | | | | | | | | | |
| Label 1 | ▲ | ▲ | – | ▲ | ▲■ | ▲ | ▲ | – | ▲ |
| Label 2 | ■ | – | ■ | ■ | ▲■ | – | ■ | ■ | – |
| CV + RFID | | | | | | | | | |
| Label 1 | ▲ | ▲ | ▲ | ▲ | ▲ | ▲ | ▲ | ▲ | ▲ |
| Label 2 | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ | ■ |

(f)

Figure 2: CV and RFID supplementing each other: (a) Same colored object tracks with label assignments in (b). (c) Different colored object tracks with label assignments in (d). (e) Considering visual occlusion and intermittent RFID, different colored object tracks with label assignment in (f)

to be scattered along the true path due to propagating location errors and distortions in a 3D space. CV and RFID location accuracies are shown with circles. The inner circle around every point shows the localization error of CV and the outer circle represents that of RFID. Figures 2a and 2b shows case I where both the objects are of the same color. Figure 2a shows the object tracks and Figure 2b represents label assignments. The CV system can correctly assign labels to ▲ as label *1* and ■ as label *2* up until time frame $t = 4$. Thereafter, there is a probability of no identity assignment, which based on relaxation labeling means no wrong label elimination and is represented here using both labels for both points. On the other hand, RFID provides correct label assignments other than at $t = 5$ due to fully overlapping localization error of point ▲ and ■. In this case RFID helps CV to generate correct object tracks. However, no label elimination in the intersection area is possible. The only contribution of CV is that it refines location.

Figures 2c and 2d shows case II where objects are of different color. Due to no occlusion CV will be able to provide correct label assignments. However, RFID will have no label assignments at $t = 5$. Fusing both feeds, CV supplements RFID here and the label assignment at $t = 5$ is obtained. For both cases CV support can also be clearly appreciated when RFID location error is maximum (i.e, on outer circle boundary) for two points having overlapping localization error in consecutive frames.

Figures 2e and 2f showcase III where some of the objects are occluded by the background and the RFID feed is intermittent. This is represented as missing vision and/or RFID location accuracy circles. The dash symbol shows non-availability of observation for that time instance. Comparing Figure 2e and 2f it is obvious that CV and RFID supplement each other at missing spots and fusion of these generates correct label assignments.

To realize how the dynamics of relaxation labeling can fuse information we describe here some related critical test cases.

Figure 3 shows case IV with simpler dynamics where two objects having the same color features moving from west to east first converge, move side by side for some time, and then diverge. The number of possible compatible labels after fusion is shown below each block of time frames.

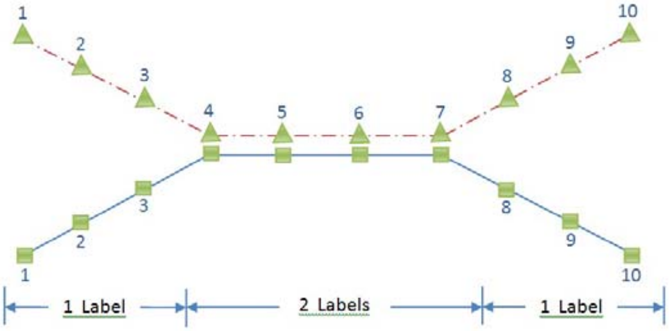


Figure 3: Correct object tracks with possible compatible labels at each block of time frames

Consider case V that is reconfigured from the scenario given in [17]. Two persons ▲,♦ (wearing distinctive clothing) carrying two balls ●,■ move towards each other and meet at the center of the test area. They then exchange the balls and backtrack to their starting positions. Both persons and both balls are tagged. To test algorithm robustness and under increased complexity we consider that the color of the balls and the persons head gear is the same. For better representation the observed 3D points shown in Figures 4a and 4b are the simulated version of the scenario over consecutive time frames.

Figure 4a represents correct trajectories of the persons and the balls. If we assume that there is no occlusion and only the stereo feed is continuously available then Figure 4b shows incorrect trajectories of the persons calculated by the stereo feed alone.

It is assumed that we have prior information about the feature set and 3D location of the object labels at time frame $t = 1$ and *2*. For subsequent time frames we correlate CV and RFID information and apply constraints in *detect*, *identify*, *locate* and *smooth* passes. Constraint based label elimination by these filtering passes update the label matrix for every observed point at every time instance. Once all impossible labels are removed and no further elimination is possible then the remaining labels are considered as compatible label/s. The labels are then passed on to the post-processing optimization process for updating fused token feature vector *v* and the refined location *XYZ* and, where required, determining a possible unique label in the compatible labels set. The labels acquired are then assigned to the observed points respectively. Figure 4c demonstrates a typical label matrix on the left that shows all four passes with the remaining compatible labels at the end. The RFID location information on the right is shown with each label matrix to provide evidence of objects presence.

For the observations in Figure 4a a step-by-step explanation on how the label matrices are updated is provided in Figure 5. For time frame $t = 3$ and *4* in each label matrix the objects are detected in the *detect* pass based on motion, color and identity and subsequently all the possible labels are assigned to all the observed object points. In the *identify* pass the system identifies objects based on color groups and identity from RFID. Since the color information for the observed points is the same, it is considered that no label is eliminated at this pass via color histogram based similarity. In the *locate* pass the labels are eliminated based on near neighbors where thresholding is done using sensor location accuracy and object speed. This helps identify ▲,● and ♦,■ as consistent label pairs. The two inconsistent labels are then eliminated from the respective label matrices. The *smooth* pass correlates labels with RFID and deletes one further label with unlikely motion according to local (*3 point* or *2 point*) smoothness and object height constraints. This would leave more global tracking to post processing after all the relaxation is completed. Note that at $t = 5$ the process of label elimination is complex due to the overlapping location errors of stereo and RFID making label elimination impossible in the *detect*, *identify* and *locate* passes. During the *smooth* pass, RFID provides no label elimination

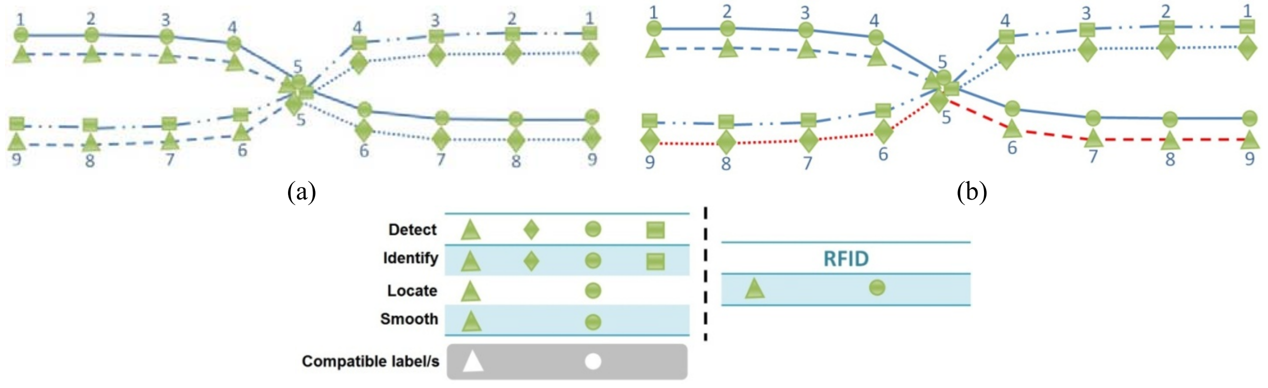


Figure 4: (a) Correct trajectories of persons and balls. (b) Correct balls and incorrect persons trajectories. (c) Left matrix - General pattern of four relaxation constraint passes and final compatible labels. Right matrix - RFID location information

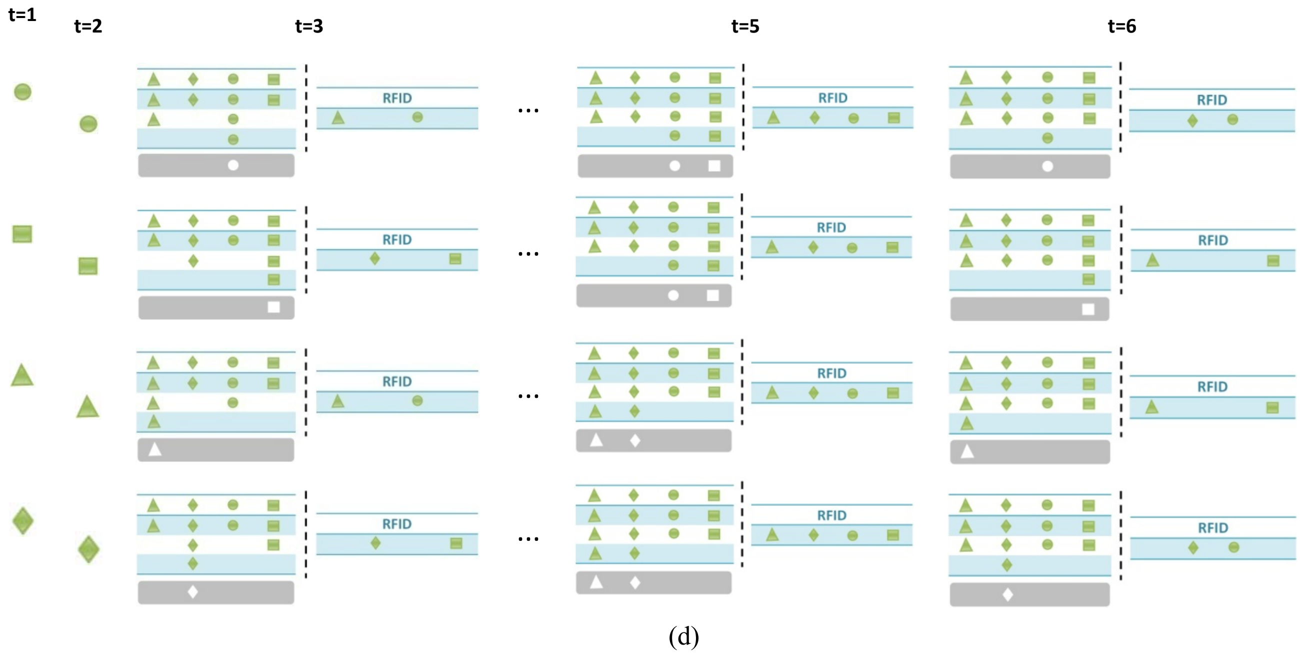


Figure 5: Label matrix updating steps for same colored objects at each time frame for Figure 4a tracks

information showing all four labels ▲,♦,●,■ as valid, however, the system identifies ●,■ and ▲,♦ as possible label set pairs based on object height and velocity constraint and subsequently outputs two compatible labels.

The compatible labels are then fed to the post optimization process to identify the optimal label for each observation. Note that the system keeps one extra label as part of the possible compatible label set. This explains a tradeoff between increased post processing computation for keeping a wrong label and the cost of eliminating a correct label. Since the objects have the same color and are assumed to be moving with the same velocity at $t = 6$ the color and near neighbor constraints will not provide information for label elimination. In the *smooth* pass based on height and direction of flow relative to the previous velocity vector direction, CV identifies ▲,●, and ♦,■ as compatible label sets for respective observations. These two label pairs for each observation represent correct trajectories of the balls, but incorrect trajectories of the persons as shown in Figure 4. However, RFID provides ♦,● and ▲,■ as possible label pairs. Correlating this information helps obtain one correct compatible label for each observation.

### 5.2 Object Color Variations

We collected various samples and analyzed HSV color space consistency for the blue and yellow balls in different weather (winter and summer) and illumination (sun and shade) conditions as shown in Figure 6.

The results shown in Figure 7 show the color consistency for

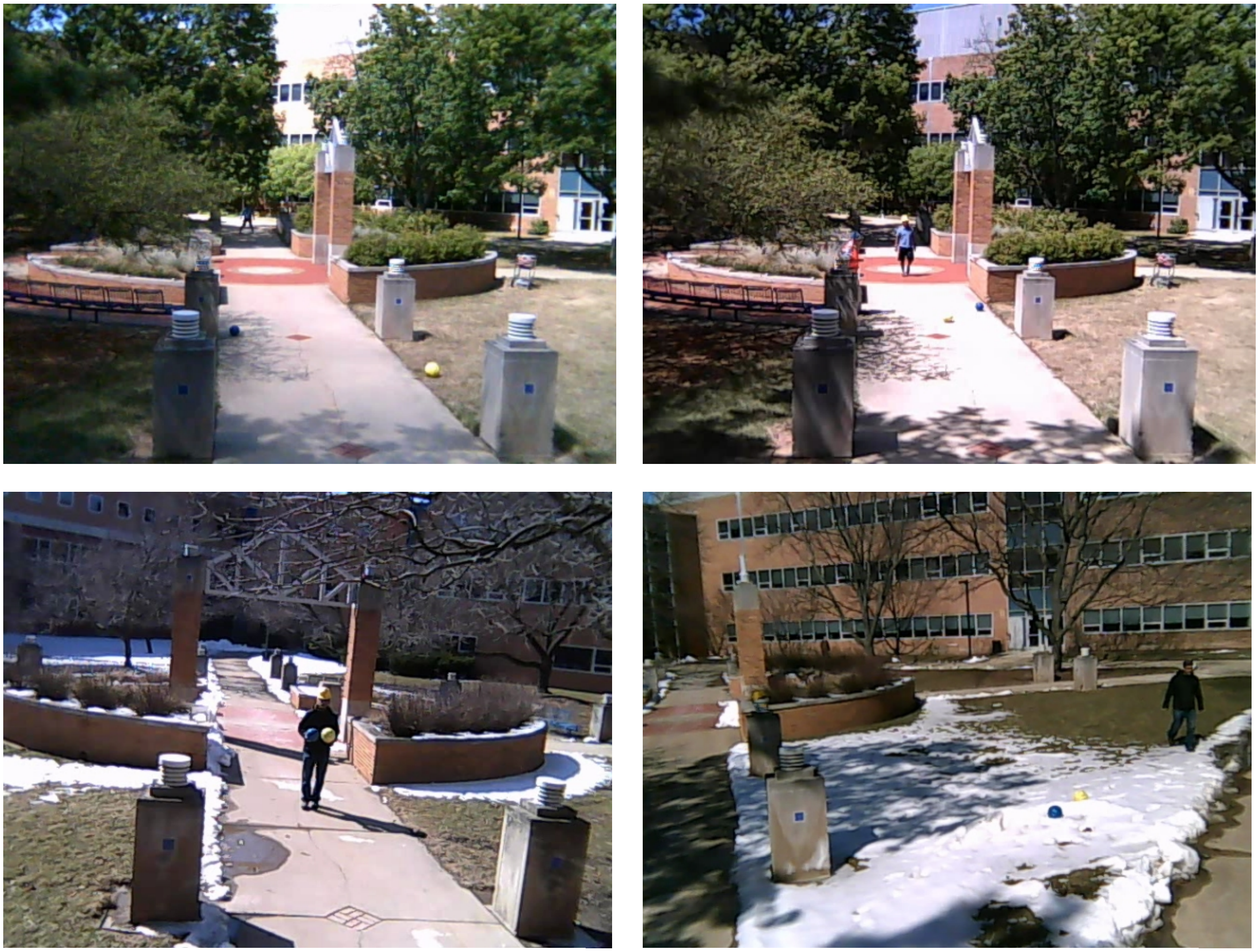


Figure 6: Different weather and illumination conditions

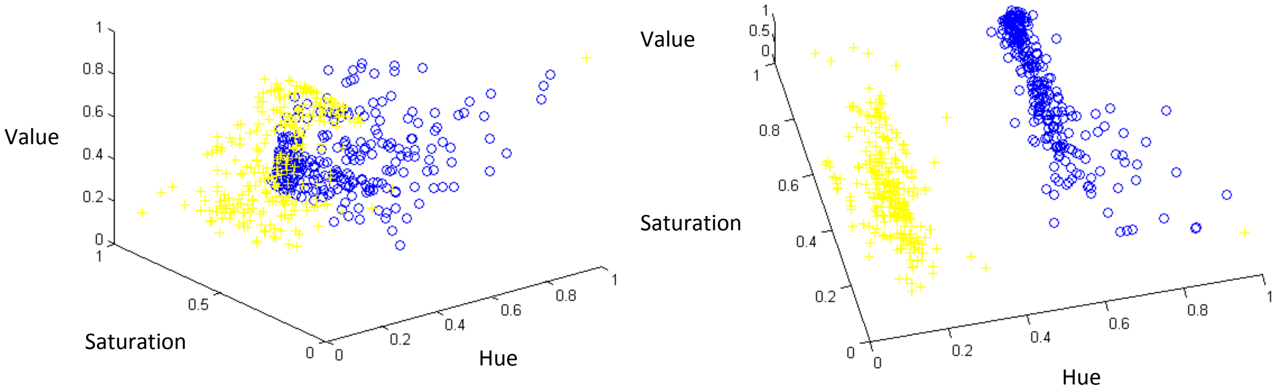


Figure 7: Analyzing blue (O) and yellow (+) ball color consistency in HSV color space under different weather and illumination conditions

blue and yellow balls for reliable color clustering. Yellow ball HSV value is represented by + and the blue ball HSV value is represented by O. The color clusters are clearly separated along the hue axis, which proves usefulness of CV to help distinguish objects based on color in an outdoor environment.

The irregular outdoor illumination variations and abrupt changes of brightness is evident in Figure 6. If color is to be used by CV to help tag and distinguish objects, then the objects must be for the most part distinguishable in the video images. In many cases workers will be wearing hard hats or vests of special coloring. The SSS should be able to take advantage of these distinctive colors by exploiting color consistency for reliable color clustering.

The experiments reported in this paper did not use automatic color similarity computations to distinguish the class of object color: instead, a symbolic color was assigned to the token.

### 5.3 Simulations of Object Tracking

Prior to collecting real outdoor data, we performed many simulations in order to assess how effective labels could be in tracking under smoothness constraints – using observations of location but not color. We created many ground truth object paths using real stereo observations made in our calibration track volume. A brightly colored ball was waved within a *69x81x61 cm* track volume and the stereo system computed the path of the ball in 3D. This was repeated ten times so that we had ten paths within the same workspace [16].

Resulting ground truth paths are shown in Figure 8. We could then take subsets of these paths for simulations. *N* observations over the time steps *1...T* were selected and presented to our tracking algorithm to see what tracks would be aggregated using the naïve physics constraints. Smoothness of trajectories requires a burst of time frames to reliably compute track smoothness, curvature, and acceleration. If we consider *n* objects and a burst of *m* time frames, then the number of possible paths will be $(n)^m$. Assume that *T* is divisible by *m*. If there is no identity information available then the number of combinations for *T* time frames will be $(T/m)X(n)^m$. Depending upon the probability *P* for an observation identity being available for the burst, the combination volume is reduced accordingly. It is considered that the identity when present is available for the whole burst. For example with *n = 3*, *m = 4* and *T = 60* the total combination volume will be *1,215* possibilities. As shown in Figure 9, with *P = 0.267* the combinations volume is reduced upto *435*.

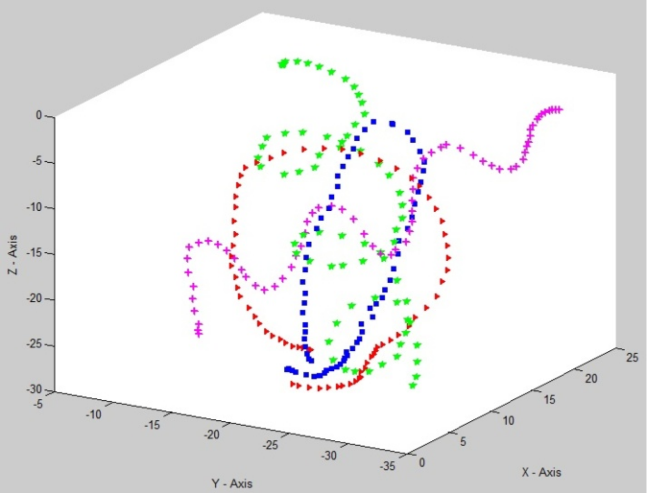


Figure 8: Ground truth trajectories generated using real stereo rig in *69x81x61cm* track volume

Simulations were conducted using *N=5, 6* and *10* object tracks and *T=60* time steps. Using probability *P*, the ground truth identity was provided in the token. Figure 10 shows results for reduction in combination volume with increase in probability *P* of object identity in the token. Computation time is also shown at marked places to realize the reduction in volume. With respect to our outdoor experiments, the probability *P* represents the time percentage for which the RFID feed for a tag was available. The algorithm was run with frame burst length *m=4*. Identity of the bursts was assumed to be randomly available. Figure 10 shows that while tracking *10* objects the combination volume can be decreased up to *99.9 percent* with the partial identity feed thereby reducing computation time.

The effect of having some identity in the tokens increases as the number of object tracks *N* increases. This data shows the difficulty faced by tracking algorithms that only use motion of image points to aggregate object tracks. Without any object identity, quantifying motion over several time steps leads to too many possible tracks. Although color, shape and texture features can be used by a passive CV system, the reliability of unique labels from RFID can yield correct tracks with far less computation. These simulations motivated us to implement an actual Site Safety System using fusion of CV and RFID.

## 6 Concluding Discussions

We have argued that the fusion of CV and RFID can produce more accurate object tracking and do so using more efficient computation. The basic reason is that RFID can provide highly reliable unique object identification, although with coarse object location, while CV can provide more accurate object location along with confirming visual features. We

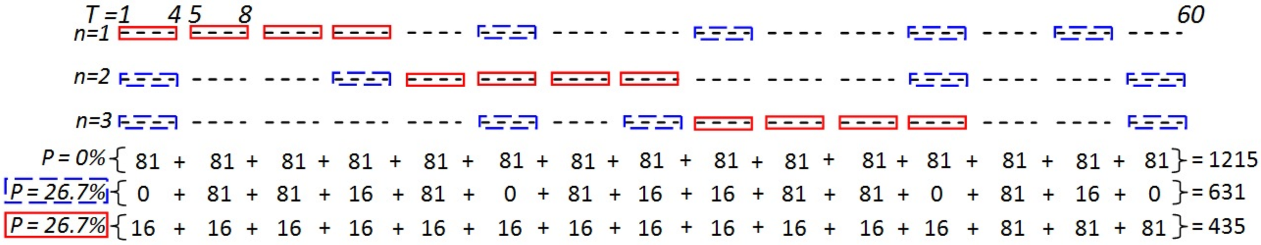

T =1 4 5 8 60
n=1
n=2
n=3
P = 0% { 81 + 81 + 81 + 81 + 81 + 81 + 81 + 81 + 81 + 81 + 81 + 81 + 81 + 81 + 81 } = 1215
P = 26.7% { 0 + 81 + 81 + 16 + 81 + 0 + 81 + 16 + 16 + 81 + 81 + 0 + 81 + 16 + 0 } = 631
P = 26.7% { 16 + 16 + 16 + 16 + 16 + 16 + 16 + 16 + 16 + 16 + 16 + 16 + 81 + 81 + 81 } = 435

Figure 9: Reduction in combinations with probability of random identity information availability

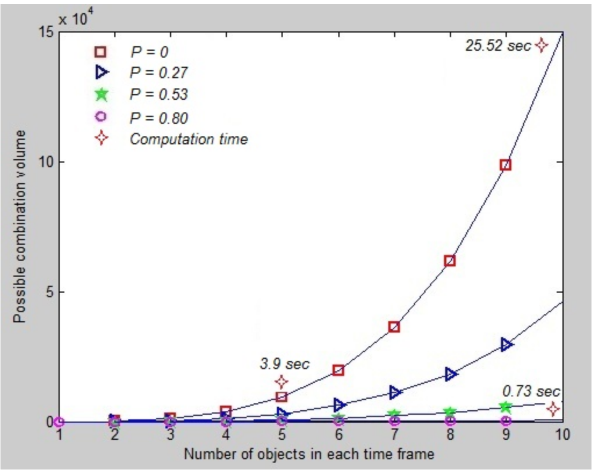


Figure 10: Possible combination volume with *n* objects and probability *p* of object identity in bursts of 4 tokens

have performed fusion experiments with tagged moving objects in a complex outdoor environment and the results support our predictions. For RFID we have used a commercially available Real Time Location System [3] and we developed our own stereo system with a laptop, MATLAB, and two commodity color cameras. Our total hardware cost was only about *US$5500*, for both the RTLS and only one stereo pair of cameras. High level performance would require more cameras and more RFID readers in the workspace than we have used. One significant problem in fusing the RFID and CV feeds is the difference in sensing frequency. Commodity cameras are designed to represent human motion well and produce upwards of *10* video images per second, whereas our RTLS system produced tokens for all tags at *2 sec* intervals. Engineering faster RFID updates will likely reduce the number of objects that can be sensed; however, this should be a favorable tradeoff in a construction site. It may also be good design to have a hierarchy of RFID sensing with a slow system for asset/material inventory and a fast system for critical objects such as workers and moving machinery.

We established that the RMS location accuracy of our stereo system is *19.3 cm* in *x, y,* and *z* for trajectories upto *24.4 m* from the cameras in a workspace that is *40x40 m*. The location accuracy for RFID was about *1.5m* in *x* and *y* ground coordinates for static objects, but about *2 m to ~ 2.6 m* for moving objects, which we attribute to the location update frequency. These results are consistent with previously published tests [9]. A commercial system should cover the workspace with a network of cameras. We have demonstrated cases where fusion disambiguates object tracks and we have also given cases where disambiguation is impossible, as in the well known shell game. We demonstrated how uncooperative objects can cheat the system. However, in general fusion of RFID and CV is better than using only one mode alone and, where costs are justified, will produce systems that are better than those using only one modality. Moreover, an automatic system detects the ambiguities and can cue the attention of higher level processes or longer lived processes, including the attention of human security personnel.

Simulations of tracking over many ground truth paths demonstrates how knowledge of unique object identity for some time instances can significantly improve correct tracking as well as reduce computation time in producing the tracks. Thus, many more objects can be tracked in practice if fused sensing is available compared to tracking by CV alone. A fast tracking implementation would be active – it could plan more efficient work, warn of possible collisions, or detect illegal operations. Finally, it is clear that the global workspace view we have used is too imprecise for detailed object interactions, such as cooperation compared to collision, or handing off carried objects. Object born touch or looming sensors would be needed for some applications. Our current work shows that pursuit of these extensions should be fruitful.

Discrete relaxation was chosen to control tracking so that we could easily experiment by switching on or off sources of information and develop our software in a modular way. Moreover, the label elimination approach easily represents the ambiguity occurring in real-life applications. The key to reducing the computational requirements is to eliminate many labels at each filtering step while keeping those labels compatible with observation. If there are *N* objects and *N* labels, the computational complexity of tracking is potentially of the order $N^{2}$ across just two time steps. We need to continue to develop our system to perform the lower level token combination and to test it fully using a set of objects with some typical behavior. We will also make the revisions that allow objects to appear and disappear from the surveyed workspace. Also much of what has been discussed assumed objects were single independently tracked points. Clearly, some objects would be a rigid aggregate of points. For example, a truck might have a single RFID tag and perhaps four or eight visual markers that would reduce combinatorics and enable rigid motion analysis. Such planar rigid structures and symmetries are also helpful to track moving objects with wide variations in position and orientation.

## References

[1] *Active RFID and Sensor Networks 2011-2021*, http://www.idtechex.com/research/ reports/active-rfid-and-sensor-networks-2011-2021-000255.asp, Accessed August 22, 2013.

[2] H. Chae and K. Han, “Combination of RFID and Vision for Mobile Robot Localization,” *Proceedings of the International Conference on Intelligent Sensors, Sensor Networks and Information Processing Conference*, pp. 75-80, Dec. 2005.

[3] *Convergence System Ltd. RTLS Development Kit*, http://www.convergence.com.hk/rtls-development-kits/, Accessed August 22, 2013.

[4] R. Cucchiara, C. Grana, M. Piccardi, and A. Prati, “Detecting Moving Objects, Ghosts, and Shadows in Video Streams,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10):1337-1342, Oct. 2003.

[5] *Dallas Zoo Tracks Elephants using CSL Real Time Location System*, http://rfid.net/news/399-dallas-zoo-

track-elephants-real-time-location-system, Accessed August 22, 2013.

[6] D. Hanny, M. Pachano, and L. Thompson, *RFID Applied*, John Wiley & Sons, Hoboken, N. J., 2007.

[7] H. Hontani, M. Nakagawa, T. Kugimiya, K. Baba, and M. Sato, "A Visual Tracking System using an RFID-Tag," *Proceedings of SICE Annual Conference*, pp. 2720-2723, Aug. 2004.

[8] *How Passports Work*, http://www.howstuffworks.com/passport.htm, Accessed August 22, 2013.

[9] *How to Install a RTLS*, http://rfid.net/basics/rtls/241-how-to-install-a-real-time-location-system-rtls, Accessed August 22, 2013.

[10] S. Jia, J. Sheng, and K. Takase, "Obstacle Recognition for a Service Mobile Robot Based on RFID with Multi Antenna and Stereo Vision," *Proceedings of the IEEE International Conference on Information and Automation*, pp. 125-130, June 2008.

[11] C. Lin, W. Peng, and Y. Tseng, "Efficient In-Network Moving Object Tracking in Wireless Sensor Networks," *IEEE Transactions on Mobile Computing*, 5(8):1044-1056, Aug. 2006.

[12] T. Meier and K. Ngan, "Automatic Segmentation of Moving Objects for Video Object Plane Generation," *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5):525-538, Sept. 1998.

[13] S. Nakagawa, K. Soh, S. Mine, and H. Saito, "Image Systems Using RFID Tag Positioning Information," *NTT Technical Review Journal*, 1(7):79-83, 2003.

[14] A. Otoom, H. Gunes, and M. Piccardi, "Feature Extraction Techniques for Abandoned Object Classification in Video Surveillance," *15th IEEE International Conference on Image Processing*, pp. 1368-1371, Oct. 2008.

[15] C. Peng, G. Shen, Y. Zhang, Y. Li, and K. Tan, "Beepbeep: A High Accuracy Acoustic Ranging System using Cots Mobile Devices," *Proc. ACM Conf. Embedded Networked Sensor Systems -SenSys*, pp. 1-14, 2007.

[16] R. Raza and G. Stockman, "Target Tracking and Surveillance by Fusing Stereo and RFID Information," *Proc. of SPIE 8392, Signal Processing, Sensor Fusion, and Target Recognition XXI*, 83921J, April 2012.

[17] R. Raza and G. Stockman, "Fusion of Stereo Vision and RFID for Site Safety," *Proc. of ISCA 25th International Conference on Computer Applications in Industry and Engineering*, Nov. 2012.

[18] I. Sethi and R. Jain, "Finding Trajectories of Feature Points in a Monocular Image Sequence," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(1):56-73, Jan. 1987.

[19] L. Shapiro and G. Stockman, *Computer Vision*, Prentice-Hall, Upper Saddle River, NJ, 2001.

[20] *Shell Game*, http://en.wikipedia.org/wiki/Shell_game, Accessed August 22, 2013.

[21] T. Teixeira, G. Dublon, and A. Savvides, "*A Survey of Human Sensing: Methods for Detecting Presence, Count, Location, Track and Identity*," ACM Computing Surveys, V, 1-35, 2010.

[22] *Triangulation and Trilateration*, http://en.wikipedia.org/wiki/Triangulation, Accessed August 22, 2013.

[23] C. Veenman, M. Reinders, and E. Backer, "Motion Tracking as a Constrained Optimization Problem," *Pattern Recognition*, 36(9):2049-2067, Sept. 2003.

[24] S. Zekavat, H. Tong, and J. Tan, "A Novel Wireless Local Positioning System for Airport (Indoor) Security," *Proc. SPIE 5403, Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense III*, pp. 522-533, Sept. 2004.

[25] Q. Zheng and R. Chellappa, "Automatic Feature Point Extraction and Tracking in Image Sequences for Arbitrary Camera Motion," *International Journal of Computer Vision*, 15(1-2):31-76, June 1995.

[26] Q. Zhou and J. Aggarwal, "Tracking and Classifying Moving Objects from Videos," *Proc. IEEE International Workshop Performance Evaluation of Tracking and Surveillance*, pp. 52-59, Dec. 2001.

**Rana Hammad Raza** received his BE Electrical from NED University and his MS degree in Computer Engineering from The Center for Advanced Studies in Engineering (CASE). He is a US Fulbright scholar and is currently working towards his PhD in the Department of Electrical and Computer Engineering at Michigan State University. He has also been working at IBM Thomas J. Watson Research Center for his graduate research. His areas of interest are object localization and tracking, site monitoring and surveillance systems.

**George Stockman** is Professor Emeritus in the Department of Computer Science and Engineering at Michigan State University, where he has been since 1982. He teaches programming and computer vision and does research in computer vision in the Lab for Pattern Recognition and Image Processing. He has been both Associate Chair and Acting Chair of the Department. He has a BS from East Stroudsburg University, an MAT from Harvard, an MS from Penn State, and a PhD in Computer Science from the University of Maryland. He is coauthor of the 2001 textbook Computer Vision with Linda Shapiro. Recent research projects include fusion of range and video for obstacle avoidance, person verification using 3D sensing, and fusion of RFID and video for work zone safety. In addition to his academic positions, he has worked four years in industry developing software for image analysis and design of automobile seating using human body and automobile models.