E-V: Efficient Visual Surveillance with Electronic Footprints

Jin Teng*, Junda Zhu†, Boying Zhang*, Dong Xuan* and Yuan F. Zheng†
*Department of Computer Science and Engineering, †Department of Electrical and Computer Engineering
The Ohio State University, Columbus, Ohio, USA, 43210
{tengj, zhangboy, xuan}@cse.ohio-state.edu, {zhuj, zheng}@ece.osu.edu

***Abstract*—Video cameras have been deployed at almost every critical location, and they keep generating huge volumes of video data. The current visual processing technologies are not efficient in handling all these data for surveillance purposes, and a large amount of human power is needed to process them. In this paper, we propose the E-V system, which uses electronic footprints to help sort through this swamp of data. Electronic footprints are wireless signals emitted by mobile devices carried by people. They are ubiquitous and amenable to collection and indexing. We study how to use electronic footprints to help *quickly* and *accurately* identify object's appearance model from large volumes of video data. We have formulated the problem and provided efficient algorithms to achieve the identification on large data sets. Real world experiments and large-scale simulations have been done, which confirms the feasibility and efficiency of the proposed algorithms.**

I. Introduction

Video based surveillance systems are deployed everywhere to continuously monitor public areas such as transport hubs, schools, government properties, etc. The purpose of such system is to identify and track objects of interest at different locations. However, video-based surveillance cannot always provide satisfactory performance due to the following challenges: (1) A lot of cameras are needed to cover a large area. The video frames generated by these cameras in large areas can easily become unmanageable over time; (2) Monitored objects may be visually occluded or have multiple inconsistent appearances.

Two typical scenarios of video surveillance are as follows: (1) Police officers track down criminals by analyzing the video sequences captured via video cameras; (2) People look for missing elders and children by searching through video sequences captured in public areas. In the first scenario, criminals may intentionally hide their faces. Therefore, police officers have to search through a large sets of video sequences to identify the appearances of the criminals. In the second scenario, people may provide an out-of-date image of the missing elder or child. Therefore, a lot of time needs to be devoted to handle the inconsistency between the missing person's different appearances. In summary, video based surveillance technologies need to involve a large amount of human efforts.

As video signals fail to provide a satisfactory solution *in isolation*, E signals emerge as a new possibility. The number of personal mobile devices being used is prolific; over four billion mobile phones are in use worldwide today [2]. The advantage of E signals is that a unique electronic identity, such as a GSM IMEI, WiFi MAC address, or Bluetooth ID, is associated with every object. If recorded, this information forms the electronic footprint of a region. Processing these E signals and footprints is generally much faster than processing video signals due to the low dimensionality of the E signal. Therefore, integrating E signals information can greatly help video based surveillance.

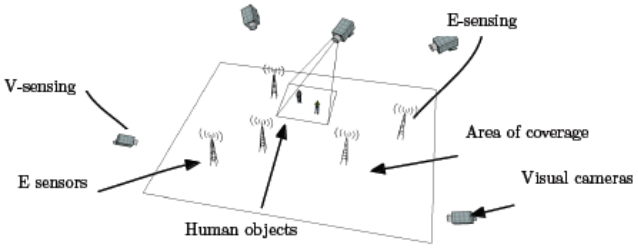


Fig. 1. E-V system

In this paper, we propose the E-V system that uses E signals to help video surveillance *quickly* and *accurately* identify object's appearance model in large areas (Fig. 1). The monitored object's visual appearance (V-ID) may be either unavailable or vague. In our system, we assume each object's E-ID is readily available. In fact, the electronic identity is mostly available, because people need to frequently use their electronic devices in daily life. Even in some extreme cases, such as the 2005 terrorism bombings in London, the E-IDs are still captured [23]. If an object's electronic identity does not exist, we leave the monitoring task to the traditional video based surveillance system.

Recall the scenarios given at the beginning. In the first scenario, police officers may not have the criminal's visual appearance, but use the criminal's electronic identity captured at the scene, with our proposed system, they may still be able to identify the criminal in recorded videos. In the second scenario, people already have a vague visual appearance of the missing person. They can use it together with the missing elder or child's electronic identity to identify the missing person in recorded videos.

There are some straightforward solutions for using E signals to help video surveillance. For example, one solution is to look up the detected E signals in a third-party supported database,

then make a matching between a particular E signal and its visual appearance. But this method may not be feasible as the E signal lookup service is not always available, or the corresponding V-ID may be vague. Another solution is online monitoring via localizing the E-ID and V-ID and mapping the locations in the two different coordinate systems. However, the E-ID localization suffers from two problems: First, E signals are inaccurate for localizing an object. E signals may attenuate significantly as distance increases. This can result in an E-ID's association with multiple V-IDs. Second, E localization is expensive. For example, we need to cover every point in a region with three or more detectors to do triangulation, which is not easy to guarantee.

In this light, we propose to use a multiple E frames' intersection approach to narrow down the search scope of a specific E-ID, EID^*, and quickly identify the corresponding V-ID, VID^*, in the corresponding V frames. We call the collection of all E-IDs and V-IDs at a certain time point an E frame and a V frame respectively. Particularly, we sort through the E-ID lists at different time points, and find a minimum set of E frames, hence corresponding V frames, where EID^* can be uniquely identified. Then, we can use the selected V frames for VID^* extraction. By taking advantage of the processing speed of E signals, our solution can greatly reduce the video processing burden. The proposed method has several favorable features: First, it is more cost friendly than those E localization based solutions. Only simple electronic detectors and video cameras are needed and they are not required to cover an area multiple times. Second, it is accurate and robust. Because we make full use of multiple frames and multiple sensors, i.e., electronic and video sensors, and perform effective information fusion, the overall accuracy is greatly boosted. Third, it is highly compatible with other approaches. It can be used to enhance or confirm the E-ID and V-ID mapping results generated by other solutions, such as online monitoring.

In summary, we have the following key contributions:

– We have proposed a methodology that leverages electronic footprints to significantly reduce the video processing time.

– We have designed a two-stage approach to find an object's visual appearance given its E-ID with or without a possibly vague V-ID. In the first stage, we find essential frames through E-ID filtering. We formulate it as a *Element Discrimination Problem* (EDP) and *Generalized Element Discrimination Problem* (GEDP). In the second stage, we identify VID^* from E filtered frames. To handle the case in which some objects' V-IDs are vague, we formulate a *n-partite Graph Best Match Problem* (nBM) and solve it via the maximum likelihood approach.

– We have conducted real world experiments and large scale simulations to evaluate the performance of the proposed E-V combination methodology. The results show that it is practical and efficient.

The remainder of the paper is organized as follows. Section II presents the related work. Section III introduces the proposed E-V system in detail. Section IV presents our experiment and simulation results. Finally Section V concludes the paper and presents our future work.

II. Related Work

There are three categories of related research work.

The first category is wireless tracking. Wireless tracking is a part of the wireless localization technology. However, wireless tracking requires the wireless network, instead of the mobile nodes, to perform the localization. So it is mainly studied under the name of Network-Based Localization [20] or Network-Centric Localization [12]. It is, in fact, implemented in almost all cellular networks and some WLAN networks. The government or police rely heavily on the location information thus obtained in some critical situations, e.g., 911 call localization or suspect tracking. In [20], Sayed et al. have surveyed the approaches popularly used and summarized the challenges faced for network-based localization. In wireless sensor networks (WSNs), localization using wireless signal emitted from one of the sensor nodes is widely studied. A good overview of this topic can be found in [18] [22] [24]. RFID tracking [19] [25] is another important part of wireless tracking. RFID is used to track children in [3] and construction material in [14].

The second category is video surveillance. Yilmaz et al. give an overview of this field in [27]. Recently there is much interest in applying a-priori constraint in object tracking [17] [15]. Multi-camera video surveillance is studied in detail in [1] [6]. A branch of video surveillance research particularly of interest is identity management [11]. The identity management problem aims to solve the continuous tracking problem after two human figures cross each other. However, the problem is confined to a homogeneous sensor network, e.g., camera networks. Integration of different types of sensors is not studied under this topic.

The last category is sensor fusion. Sensor fusion is the integration of sensing data from multiple sensors, and thereby enhances the quality of acquired information for a sensing area [10]. It has many real applications in a variety of different areas, such as robotics or bioinformatics. There are several classic theories to perform sensor fusion, such as Kalman Filtering [21] and Dempster-Shafer theory [26]. Recently there are also works focusing on using heterogeneous sensors to help tracking [16] [30]. However, the identity management problem has not been seriously addressed. It is often implied in literature that the mapping is known or can be easily done with some a-priori knowledge. Cho et al. mentioned this problem in [7]. Cho et al. use a visual camera network and a RFID network for monitoring. They propose to use the appearance or disappearance of RFIDs and human figures to perform the mapping. However, they only give an example of how the process is done. The feasibility of such an approach is not analyzed. Nor did they give an efficient algorithm for identity mapping when the volume of sensing data is huge.

III. E-V Design

There are two major steps in the identification scheme for our E-V system. First, we have an E frame filtering step, which helps to remove a large number of irrelevant E frames, hence corresponding V frames, captured in given places and retains only a small set of E frames to help identify the object of interest. Second, we have a V frame mapping step, which uniquely identifies the appearance of the object of interest. For the possible vagueness of V-IDs, this step also provides functionalities to eliminate certain levels of uncertainty. In the following two subsections, we will discuss these two steps in detail.

A. E Frame Filtering

We study two cases when E-IDs are complete and incomplete. By "complete", we mean there is no missing E-IDs or extra E-IDs (ghost E-IDs) captured in any E frame.

	EID*	EID1	EID2	EID3	EID4
e1	1	1	0	0	1
e2	1	0	1	1	1
e3	0	0	1	1	0
e4	0	1	0	1	0

(a)

	EID*	EID1	EID2	EID3	EID4
e1'	1	1	0	0	1
e2'	1	0	1	1	1
e3'	1	1	0	0	1
e4'	1	0	1	0	1

(b)

Fig. 2. Sample matrix of E and E' in (a) and (b) respectively

1) E Frame Filtering with Complete E-IDs: As we mentioned above, if we want to find a VID^* given EID^*, we may use E-IDs as indices to select essential V frames for further processing. In these frames, we must ensure that EID^* and VID^* can be uniquely identified, i.e., being 'distinguishable' from other IDs. So here we need to define 'distinguishability' of an E-ID or V-ID in a set of E frames or V frames.

We collect the E-IDs which have ever appeared across all E frames. Each E-ID is numbered and denoted as EID_{number}. EID_0 is given to EID^*. We can construct a matrix $E = \{e_{ij}\} = (e_1, e_2, \cdots, e_n)^T$, as is shown in Fig. 2(a). Its columns are EID_{number}. Its rows are different frames, denoted as e_i. e_{ij} is the j-th element in e_i. We put a 1 in e_{ij}, if EID_j appears in frame e_i, and 0 if not. Then we have Definition 1.

Definition 1. *EID^* is distinguishable from $EID_i \neq EID^*$ in an E frame set P, if and only if $\exists e_j \in P$: $e_{j0} \neq e_{ji}$.*

Definition 1 captures the fact that any $EID_i \neq EID^*$ can be distinguished from EID^*, so long as EID^* and EID_i do not always appear or disappear together in V and E frames. From here, we can have Definition 2.

Definition 2. *EID^* is distinguishable in an E frame set P, if and only if for $\forall EID_i \neq EID^*$, EID^* is distinguishable from EID_i.*

We give an example in Fig. 2(a). In Fig. 2(a), EID^* can be distinguished from EID_1 with frame e_1 and e_2. But EID^* cannot be distinguished from EID_4, because we cannot find such e_j. So EID^* is not distinguishable in Fig. 2(a). However, it is distinguishable, if we take away the EID_4 column. One note here is that we must make sure that EID^* appears at least once. That means one of the rows selected from E must have a 1 for the EID^* column. If EID^* does not appear at all, it is meaningless to talk about its distinguishability.

Definition 2 gives a general way to decide the distinguishability of an E-ID. If there is no column in E which is identical with that of this E-ID, it can be uniquely identified. Though Definition 1 is intuitive, by a slight change in elements, we can greatly extend its compatibility, as we will see later in this section.

We further define the complement operation for the matrix and vector, $\overline{E} = (\overline{e_1}, \overline{e_2}, \cdots, \overline{e_n})^T = \{1 - e_{id}\}$. We also define the conjunction operation for row vectors in E, $e_i \cap e_j = (e_{i1} \wedge e_{j1}, e_{i2} \wedge e_{j2}, \cdots)$. Now we define another matrix $E' = \{e'_{ij}\}$. If EID^* appears in e_i, then $e'_i = e_i$. If EID^* does not appear in e_i, then $e'_i = \overline{e_i}$. Transformation of the sample matrix E into E' is shown in Fig. 2. We let e^* be a vector where only the element in the EID^* column is 1, and all other elements are 0. With these definitions, we can have Theorem 1. The proof is not difficult, so we skip it due to space limitation.

Theorem 1. *EID^* is distinguishable in E, if and only if $\exists P = \{e'_{p_1}, e'_{p_2} \cdots, e'_{p_l}\}$, whose elements are all row vectors in E', such that $\cap_{i=1}^{l} e'_{p_i} = e^*$, where $e^* = (1, 0, 0, \cdots)$.*

Fig. 2(b) shows the concept of Theorem 1. With this theorem, we can formulate in Definition 3 the problem of finding a minimum set where EID^* is distinguishable.

Definition 3. *Element Discrimination Problem (EDP): Find a minimum set (in term of element number) $P = \{e'_{p_1}, e'_{p_2} \cdots, e'_{p_l}\}$, whose elements are all row vectors in E', such that $\cap_{i=1}^{l} e'_{p_i} = e^*$.*

Theorem 2. *EDP is NP-Complete.*

We sketch the proof here. In EDP, we need at least one 0 along each column in the selected P. It means that all 0s in E' must cover every column except EID^*. So there is a one to one correspondence between EDP and the set cover problem, which is NP-complete.

For the set cover problem, we have a $(1 + \log(n))$-approximation scheme, where n is the number of element in the set to be covered [8]. It is shown in Algorithm 1.

After the E frame filtering step, we can use the one-to-one E-V frame mapping to find the V frames corresponding to the selected E frames and conduct visual detection techniques to uniquely identify VID^*. This will be discussed in detail in the next subsection.

2) E Frame Filtering with Incomplete E-IDs: In the previous case, E-IDs are treated as complete. However, it is often not realistic in real world systems. There are mistakes in the sensing data. It is likely that we will have false positives and false negatives, i.e., missing IDs and ghost IDs. These mistakes

Algorithm 1 Greedy_EDP(E,e^*)

$X \leftarrow$ {all VID_is which correspond to a 0 in e^*}
F $\leftarrow \emptyset$
for each row e in E **do**
  $S \leftarrow$ {all VID_is which correspond to a 0 in e}
  F $\leftarrow$ **F** $\cup$ {S}
end for
$U \leftarrow X$, **C** $\leftarrow \emptyset$
while $U \neq \emptyset$ **do**
  select an $S \in$ **F** that maximizes $|S \cap U|$
  $U \leftarrow U - S$, **C** $\leftarrow$ **C** $\cup \{S\}$
end while
return **C**

can be classified into two types. The first type is aleatory ones. It happens mainly because sensors are not perfect, and the processing of data can also yield such mistakes. For example, we miss an E-ID in an E frame, because the signals were interfered with and happened not to be received. The second type of mistakes is consistent mistakes. Some factors uncontrollable on the sensor side consistently generate false positives or false negatives. For example, in the E-V system, not everyone has an electronic device or, even if they have, they may choose not to expose their E-IDs. So a person may be seen in the video and is assigned a V-ID, but there is no corresponding E-ID detected by the wireless detectors. The V-ID in the video then becomes a ghost ID, or, equivalently, the E-ID is missing. In the following part, we will investigate how to handle these two types of mistakes.

Handling Aleatory Mistakes: Aleatory mistakes come from the imperfectness of sensors and data processing algorithms. We can definitely improve and fine tune the sensors and algorithms, but this type of mistakes will be there anyway. So it is necessary to find a generic solution to this problem. Generally speaking, we can apply pre-processing to tackle this type of mistakes. It is possible to leverage some intrinsic properties of object movements or other a-priori knowledge to filter away obviously erroneous data.

One preprocessing method to handle aleatory mistakes is smoothing. It uses a smoothing filter to do away with too abrupt changes in the sensing data (Fig. 3). It generally makes sense, as the underlying physical motion of objects is always continuous and does not see many abrupt changes. For example, human cannot teleport. So an E-ID detected is likely to re-appear in the next frame.

The smoothing can be implemented by applying a smoothing kernel to the appearance/disappearance matrix, such as E in the last section. How to choose which type of smoothing and with what parameters is case specific, and is beyond the scope of this paper. But we have observed that using a simple moving average filter along the column of E performs well in our experiments.

However, smoothing can cause problems. The most salient one is that the E matrix is no longer binary, i.e., 0/1. So we need to recast the EDP problem. Here we give the definition

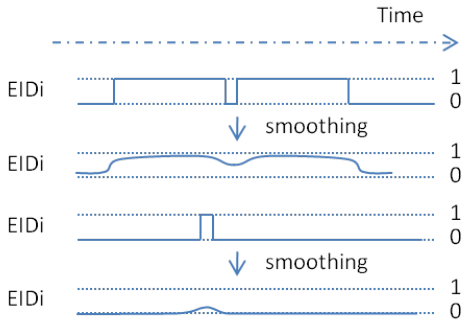


Fig. 3. Smoothing to get rid of abrupt appearance or disappearance of IDs.

of GEDP in Definition 4.

Definition 4. *Generalized Element Discrimination Problem (GEDP): Find a minimum set* $P = \{e'_{p_1}, e'_{p_2} \cdots, e'_{p_l}\}$, *such that* $d = |\cap_{i=1}^{l} e'_{p_i} - e^*| \leq \tau$, *where* τ *is a preset threshold.*

The operations are defined based on the fuzzy set theory [29], as it is the natural extension of the traditional (crisp) set theory. For example, the conjunction operation can be defined as $a \wedge b = \min(a, b)$, and disjunction as $a \vee b = \max(a, b)$. The distance between two vectors, i.e., $|e_a - e_b|$, can be either Manhattan or Euclidean distance. Fig. 4 gives an example.

	EID*	EID1	EID2	EID3	EID4
e1'	0.98	0.95	0.1	0.01	0.06
e2'	0.9	0.01	1	0.94	0.04
e3'	0.88	0.99	0.03	0.1	0.12
e4'	0.99	0.02	0.89	0.27	0.23
Conjunction	↓	↓	↓	↓	↓
e	**0.88**	**0.01**	**0.03**	**0.01**	**0.04**
e*	**1**	**0**	**0**	**0**	**0**

Distance between e and e*:
Manhattan—0.21 Euclidean—0.13

Fig. 4. Sample GEDP problem

GEDP is clearly NP-hard, as we can reduce EDP to GEDP. In fact, it is an NP-hard problem to calculate d_{min}, the minimum d given all the rows in E'. As a reference to determine τ, it is very helpful to know d_{min}. Fortunately, we have an additive Fully Polynomial-Time Approximation Scheme (FPTAS), which is based on subset-sum approximation algorithms [8], that can ensure that $\tilde{d}$, the returned minimum value of d, is in the range $(d_{min} - \epsilon, d_{min} + \epsilon)$. In Algorithm 2, we show the algorithm to achieve that, as well as the heuristic solution to the GEDP problem. Here we define an operation on a list of numbers. Suppose L is a list of number, and $L \wedge x = \{l \wedge x | l \in L\}$. For example, $L = \{l_1, l_2, l_3\}$, and $L \wedge x = \{l_1 \wedge x, l_2 \wedge x, l_3 \wedge x\}$. s_i denotes the i-th element in a

set S. The Merge-Lists function merges the two lists, and sort the merged list according to the distance. The Trim function only keeps 'representative' entries in the list to keep the list of length polynomial to the frame number. In real practice, we often cut the tail of the list. We throw away entries far away from the objective. The details of both functions can be found in [8].

Algorithm 2 Approx_GEDP(E,ϵ,τ)

```
S ← {all row vectors in E}
n ← |S|, L_0 ← inf
for i ← 1 to n do
    L_i ← Merge-Lists(L_{i-1}, L_{i-1} ∧ s_i)
    L_i ← Trim(L_i, ε/2n)
end for
d̃ ← the smallest value in L_n
if d̃ > τ then
    No solution for the given τ
else
    return z* in L_n, with the minimum ∧ operands
end if
```

Besides smoothing, we may also take advantage of a-priori knowledge to help reduce aleatory mistakes. The object lists acquired may come with some measures of the object. For example, in wireless sensing, besides the MAC address, we may also have the Received Signal Strength Indication (RSSI). These measures are where we can apply our a-priori knowledge to help the mapping process.

In the E-V system, the RSSI can be used to estimate the distance of the object to the wireless detector. As the video cameras cover the sensing area of wireless detectors, the wireless detectors are likely to be near the center of the camera's range. So being closer to the wireless detector means a better chance that it is not a ghost ID, i.e., the video camera should detect the person with the device. So we can write the probability of a person's appearance in a V frame as a function of the RSSI as a sigmoid function, $f(t) = 1/(1 + e^{\beta(t_0 - t)})$. We also know through experiments or empirical formulas that RSSI statistically attenuates versus distance in an exponential manner described by the classical formula $P = P_0 - \alpha \cdot \log \frac{d}{d_0} + N$. What we need to determine now is α, d_0. and find their relationship with β and t_0. Empirically the RSSI of handsets which is detected by a WiFi AP and is also in the camera range falls into the range of -55 dBm and -75 dBm. So t_0 can be set to -65 dBm, and $\beta = 0.1 - 0.5$ gives a fairly good performance. So we can fit our sigmoid function accordingly. Again, after distance estimation, we must use the GEDP algorithm to find the essential V frames, as the elements in E are no longer 0/1.

Handling Consistent Mistakes of Missing E-IDs: Different from aleatory mistakes, consistent mistakes of missing EIDs are out of control of the system. When this is the case, it is usually very helpful to put constraints on the sensing pattern and sensor deployment. We may require relationship between certain types of ID detection. For example, we can deploy the E-V system, so that whenever an E-ID is detected, its corresponding V-ID is also detected. It can be done by making sure that the coverage of camera network is larger than the wireless coverage. This is a reasonable constraint. If an E-ID is detected, the corresponding device is very likely to be there, and the video camera should detect it. With this constraint, we can still use the identity mapping approaches presented in the last subsection. But the EDP algorithm is more limited in this case, only E frames with EID^* appearing make sense. So we can only make use of those frames to distinguish EID^*. However, we do need some frames where EID^* is guaranteed not to be there, in order to eliminate the environmental E-IDs, as they are always present. It can be done by selecting videos recorded at a time when nobody is there, e.g., during the dead of night.

As the sensor deployment and sensing pattern are at full discretion of the sensor network designer, the above method is almost always possible and of generic utility. We also emphasize that exacting constraints on sensing patterns and sensor deployment is not so unreasonable as it may sound. After all, if all sorts of IDs can randomly be missing or become ghost IDs in a consistent manner, it is not possible at all to perform identity mapping.

B. VID^* Identification from E Filtered Frames

In this subsection, we discuss our scheme for identifying VID^* from E filtered frames. Once the EDP/GEDP algorithms have selected a set of frames that can distinguish EID^*, VID^* is guaranteed to be identifiable from these frames. We will first discuss the cases where no (vague) V-ID input has been given. We will address how to integrate the V-ID input at the end of this subsection.

Since the observed feature of an object varies slightly due to the uncertainty in pose and/or illumination changes, the V-IDs related to the same object are similar to each other but usually not identical. We set up a similarity matrix between any V-IDs, from the same frame or different V frames, as is shown in Fig. 5. We write the similarity between VID_i and VID_j as $s_{(VID_a, VID_b)}$ or s_{ab} when there is no confusion. It is symmetric. We fix the similarity between two V-IDs appearing in the same V frame to be 0, because they cannot belong to the same person. Since it is desirable to pinpoint a person or an E-ID within a short period of time, e.g., within one day, we assume that the association between E-IDs and V-IDs generally do not change. Even if people occasionally change certain types of appearances, e.g., clothing, yet it is still possible to differentiate between different people based on other features, e.g., gaits [13], and have a reasonable similarity metric.

The similarity between two V-IDs provides a probabilistic measure about whether they come from the same object. If we compare the same person in different frames, we may get a distribution of similarity. We call this distribution function $f_1(s)$, as is shown in Fig. 6(a). If we compare all pairs of different people across all frames, we may also get a

	v1		v2			vn	
	VID1	VID2	VID3	VID4	...	VID(m-1)	VIDm
VID1	N/A	0					
VID2	0	N/A					
VID3			N/A	0			
VID4			0	N/A			
...					...		
VID(m-1)						N/A	0
VIDm						0	N/A

Fig. 5. Similarity Matrix

distribution of similarity. We call this distribution function $f_2(s)$. These two similarity functions reflect how likely two V-IDs are from the same person or different people. For a VID_i and VID_j, the probability of their belonging to the same person is $P(VID_i = VID_j) = f_1(s_{ij})$, the probability of their belonging to different people is $P(VID_i \neq VID_j) = f_2(s_{ij})$. However, these two functions are hard to get and they may change in different scenarios. So we approximate them with $f_1'(s)$ and $f_2'(s)$ as is shown in Fig. 6(b). It means the larger similarity, the more likely the two V-IDs belong to the same person. It is reasonable in common sense. Then $P(VID_i = VID_j) \approx s_{ij}$, $P(VID_i \neq VID_j) \approx 1 - s_{ij}$. Given a set of V-IDs, $\{VID_1, \cdots, VID_k\}$, the probability that VID_0 is none of them is

$$P(VID_0 \neq VID_i,\ i = 1, \cdots, k) = \Pi_{j=1}^{k} f_2(s_{0j}). \quad (1)$$

The probability that VID_0 is $VID_i (1 \leq i \leq k)$ is

$$P(VID_0 = VID_i) = \frac{f_1(s_{0i})}{f_2(s_{0i})} \cdot \Pi_{j=1}^{k} f_2(s_{0j}). \quad (2)$$

Note that the above probability should be normalized to exclude the probability that VID_0 equals VID_i and VID_j from the same V frame at the same time. In favorable conditions, many terms in the product in (1) and (2) are close to one or to zero, so we can further simplify the probability that VID_0 is none of them to $1 - \max(s_{0i})$, and that VID_0 is one of them to $\max(s_{0i})$. These two approximations are much simpler to calculate and they work well in real experiments.

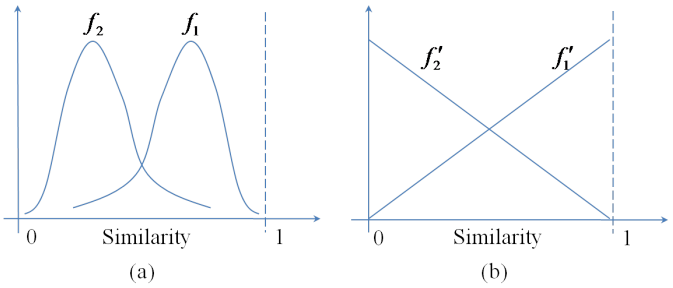


Fig. 6. Similarity distribution

The appearance/disappearance probabilities of V-IDs based on the similarities can be matched with the appearance pattern of EID^* to identify the candidates for VID^*. Suppose if we have n V frames selected by the EDP/GEDP algorithm, named v_1 to v_n. And there are m_i $VIDs$ in v_i, named VID_1^i to $VID_{m_i}^i$, which is shown in Fig. 7(a). Then we form the problem of finding VID^* as in Defintion 5.

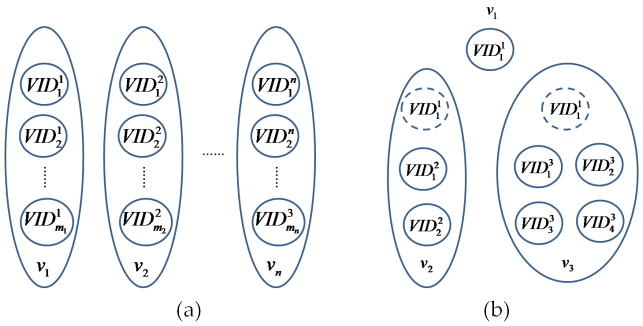


Fig. 7. n-partite graph and the n-partite Graph Best Match Problem

Definition 5. *n-partite Graph Best Match Problem (nBM): Find a V-ID, the product of whose appearance/disappearance probabilities matches that of EID^* best.*

The intuition behind Definition 5 is that we know only VID^* appears or disappears in the same way as EID^*. So, to solve the nBM problem, we can evaluate the appearance or disappearance probability of each V-ID in every frame. Let us take an example (Fig. 7(b)). Suppose we know that EID^* appears in V frame v_1 and v_3. To choose which V-ID is VID^*, we calculate an appearance/disappearance probability for each V-ID. For example, we first pick VID_1^1 for evaluation. We try to figure out whether it has appeared in each V frame with (1) and (2),

$$\begin{aligned} &P(VID_i = VID^*) \\ &= P(VID_i \in v_1) \cdot P(VID_i \notin v_2) \cdot P(VID_i \in v_3) \end{aligned} \quad (3)$$

We can do the same for every V-ID. After that, we choose the V-ID with the largest probability to be VID^*, because its appearance/disappearance tallies with that of EID^* best.

Finally, let us discuss what if a (vague) V-ID input has been provided. With the input V-ID, we can calculate a similarity score between the input V-ID and each V-ID we have obtained from the image. Then we incorporate this score into the matrix shown in Fig. 2. One typical way is to calculate a weighted average of the new similarity score and the original similarity scores in the similarity matrix. If we are very confident that the input V-ID is accurate, we can assign a larger weight to the new similarity score. Otherwise, we need to assign a smaller weight.

Remark: We would like to point out that, the purpose of filtering E frames is to reduce the visual processing burden. However, if the visual processing capability is not very limited, we can keep more E frames, e.g., twice as many as minimal. It will give better accuracy for VID^* identification.

IV. Evaluation

In this section, we comprehensively evaluate the proposed E-V system. We have done two real world experiments and a set of large-scale simulations. We will present the results below.

A. Real World Experiments

We have done two experiments to validate the identity mapping schemes designed for the E-V system. One experiment is done in the gymnasium, and the other is in the library. Each experiment lasts several minutes. Aside from environment E-IDs and those of passers-by, we have six colleagues carrying E-IDs in the first experiment, and eight in the second experiment. We are able to extract 28 corresponding E and V frames for the first experiment and 40 for the second. In both experiments, we set up a camera shooting from above and covering the entire area. V-IDs are detected using the HoG pedestrian detector [9] and their similarities are calculated from the color histograms. We also put a laptop on the ground to record WiFi MACs, it is close to the center of camera scope.

Our purpose is to find a specific VID^* given an EID^*. We have applied all the techniques introduced above, e.g., E frame filtering, maximum likelihood V-ID selection, deployment adjustment, smoothing and distance estimation. In both experiments, we have successfully found the VID^*. Some sample V frames in the final outcome are shown in Figs. 8 and 9. VID^* is circled with the green box. We will report some of the details below.

	Optimal	EDP	GEDP(tight)	GEDP(loose)
Experiment 1	3	5	4	3
Experiment 2	3	6	7	4

TABLE I
Number of V frames selected

We have tried EDP and GEDP to find the minimum number of V frames for further processing (Table I). In both experiments, the optimal number is 3. With EDP, we are able to select 5 and 6 E frames to distinguish EID^*. With GEDP, we set two τ values, with $\epsilon = 0.3$. One is very close to the minimum distance achievable. The other is that distance plus 2. The minimum distance achievable is around 0.6 for the first case, and 2.3 for the second. They are remarkably low, as we have over 100 E-IDs and the largest possible distance can be over 100 (we use Manhattan distance). With tight threshold, i.e., smaller τ, we have selected 4 V frames for the first experiment, and 7 for the second. And the loose threshold gives 3 and 4 V frames respectively. All the results can uniquely identify EID^*, except the GEDP(loose) for Experiment 1. There is one E-ID that cannot be told apart from EID^*. However, those two E-IDs are very different visually, so it does not affect the final outcome.

B. Large-Scale Simulations

To further validate the performance of our proposed identity mapping scheme for the E-V system, we have also performed multiple large-scale simulations. In our simulations, there are a total of 120 objects each associated with a given E-ID and original image. The original image is chosen from the pedestrian samples from INRIA person database [9]. The V-ID similarity is calculated from the color histograms same as the real world experiments. By using random waypoint model, all objects are distributed across 4 separate scenes each with an area of 100×100 m. Depending on the coordinates of a given object within the corresponding scene, its RSSI as received by the E detector is simulated, and a perspective distortion determined from its distance to the camera is applied to its original image corrupted by randomly generated image noises to simulate its V-ID.

We measure the performance of our proposed scheme on two aspects: 1) the efficiency for E frame selection, which is measured by the percentages of output frames from EDP and GEDP algorithms with respect to the total input frames; and 2) the accuracy for identifying VID^*. As can be seen from Fig 10, both EDP and GEDP algorithms greatly reduce the number of frames for further V processing. The percentages of selected V frames decreases as the numbers of input frames increase, which demonstrates the scalability of our proposed scheme when a large number of input frames are provided. For the GEDP algorithm, selecting τ larger than 4 yields a smaller number of output frames.

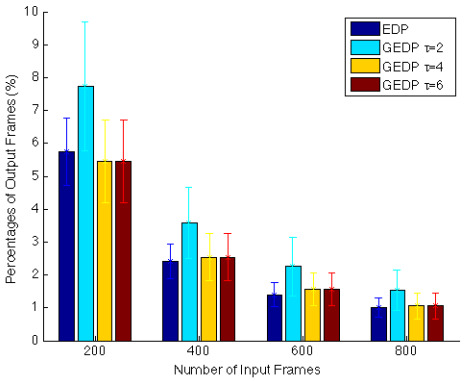
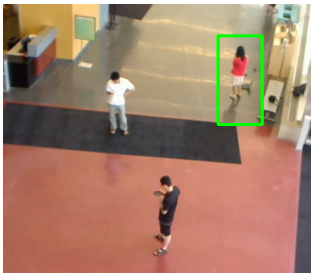


Fig. 10. Percentages of selected V frames with respect to the total frames.

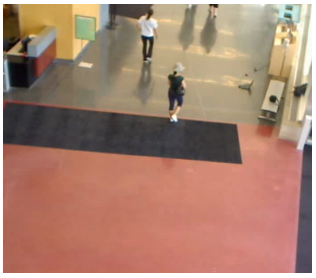
Fig. 11 shows the results on the accuracy rates for identifying VID^*. Over multiple repeated runs on random generated E and V frames, the average accuracy rates are over 95% for EDP and GEDP with parameter settings. Different numbers of input frames do not have much impacts on the accuracy for identifying VID^*.

In Fig. 12, we simulate the V-ID detection failure rate under different E-ID missing rate at 1%, 10%, 30% and 50%. Generally the failure rate rises when the E-ID missing rate increases. But even when the missing rate is as high as 50%, the detection rate is pretty good at around 90%. Considering the practical issues on the V side which may cause a V-ID not being detected in certain scenarios, we have also taken the miss detection into consideration and studied its impact on the identity mapping scheme. A V-ID is randomly discarded

(a) Sample Frame 1

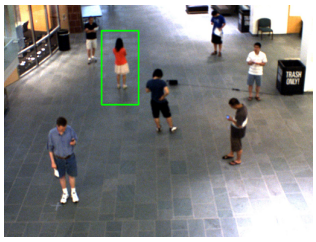


(b) Sample Frame 2

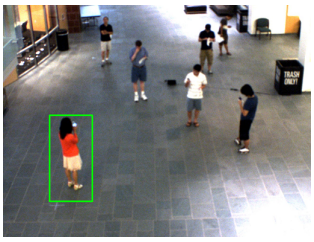


(c) Sample Frame 3

Fig. 8. Experiment One



(a) Sample Frame 1



(b) Sample Frame 2

(c) Sample Frame 3

Fig. 9. Experiment Two

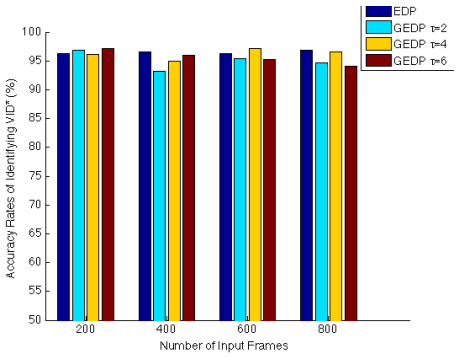


Fig. 11. Accuracy rates for identifying VID^*.

according to the specified miss detection rate to simulate the miss detection in practical vision systems. Different miss detection rates at 2.5%, 5%, 7.5% and 10% are simulated and the results are shown in Fig. 13. The deterioration of V detection does have negative impacts on the accuracy rates. However, even at a very high miss detection rate of 10% which is far below the state-of-art person detection performance, our V-ID matching scheme is still able to provide reasonable accuracy rates above around 70% over the selected frames from both the EDP and GEDP algorithms.

V. Final Remarks

In this paper, we have introduced the E-V system. Electronic footprints and visual images are integrated in the E-V system to facilitate data processing. The E-V system is essentially a bi-modal sensor fusion network. We have addressed the problem of efficiently identifying object's visual appearance with the help of electronic footprints from large volumes of video data. We formulated the problem, and provided efficient solutions. In the problem, we took practical situations, e.g., missing IDs and ghost IDs, into consideration, and devised schemes to eliminate their impacts. Real world experiments and large-scale simulations have been done for the E-V system. The results confirm the feasibility and efficiency of the theories and algorithms that we have developed.

As part of our future work, we willl study how to economically deploy visual and electronic sensors to fully cover an area and satisfy the constraints in Section III. As it is a problem to cover a 3D space and the cameras are directional, we will conduct this study based on our previous studies presented in [4] [5] [28]. We also plan to form larger networks for more comprehensive evaluation on our prototyped E-V system.

Acknowledgment

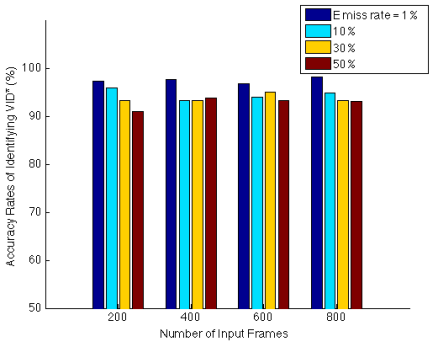
This work was supported in part by China 973 Project 2011CB302800, the National Science Foundation of China (NSFC) under grant No. 61070221, and the US National Science Foundation (NSF) under Grant No. CNS0916584, CNS1065136. Any opinions, findings, conclusions, and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

References

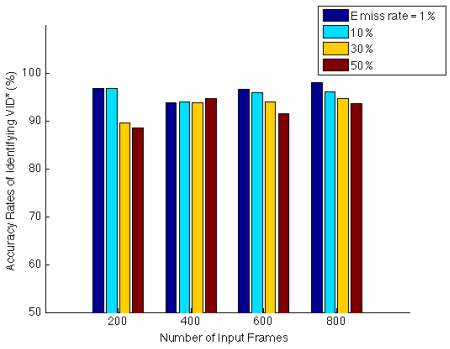
[1] H. Aghajan and A. Cavallaro, "Multi-Camera Networks", Elsevier Inc., 2009.

[2] T. Ahonen, "5-4-3-2-1, as in Billions. What do these gigantic numbers mean?", http://communities-dominate.blogs.com/brands/2010/08/5-4-3-2-1-as-in-billions-what-do-these-gigantic-numbers-mean.html, 2010.

[3] A. R. Al-Ali, F. A. Aloul, N. R. Aji et al., "Mobile RFID Tracking System", ICTTA 2008, April 2008.

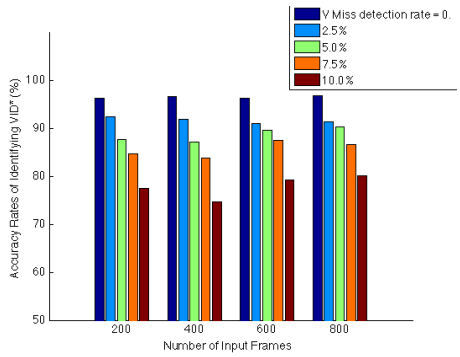


(a) EDP

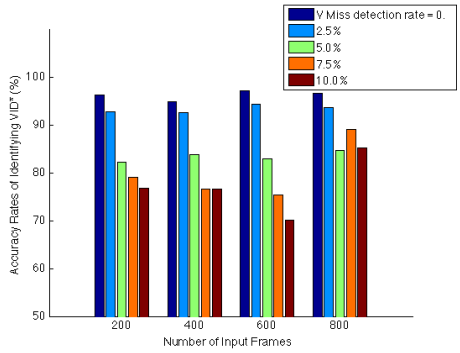


(b) GEDP ($\tau = 4$)

Fig. 12. Accuracy rates of VID^* selection considering miss detection of E-IDs.



(a) EDP



(b) GEDP ($\tau = 4$)

Fig. 13. Accuracy rates of VID^* selection considering miss detection of V-IDs.

[4] X. Bai, C. Zhang, D. Xuan and W. Jia, "Full-Coverage and k-Connectivity (k=14,6) Three Dimensional Networks", IEEE INFOCOM, 2009.

[5] X. Bai, C. Zhang, D. Xuan, J. Teng and W. Jia, "Low-Connectivity and Full-Coverage Three Dimensional Networks", ACM MobiHoc, 2009.

[6] J. Black and T. Ellis, "Multi-camera Image Tracking", Image and Vision Computing, vol. 24, issue 11, Performance Evaluation of Tracking and Surveillance, 2006.

[7] S. H. Cho, S. Hong and Y. Nam, "Association and Identification in Heterogeneous Sensors Environment with Coverage Uncertainty", IEEE Advanced Video and Signal Based Surveillance, 2009.

[8] T. Cormen, C. Leiserson, R. Rivest et al., "Introduction to Algorithms Second Edition", MIT Press, Cambridge, MA, USA, 2001.

[9] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection", in Proc. International Conference on Computer Vision & Pattern Recognition, vol. 2, pp. 886-893, June 2005.

[10] W. Elmenreich, "Sensor Fusion in Time-Triggered Systems", PhD Thesis, Vienna University of Technology, 2002.

[11] L. J. Guibas, "The Identity Management Problem - A Short Survey", International Conference on Information Fusion, pp. 1-7, July, 2008.

[12] F. Gustafsson and F. Gunnarsson, "Mobile Positioning using Wireless Networks: Possibilities and Fundamental Limitations Based on Available Wireless Network Measurements", Signal Processing Magazine, IEEE , vol. 22, no. 4, pp. 41-53, July 2005.

[13] J. Han and B. Bhanu, "Statistical Feature Fusion for Gait-based Human Recognition", CVPR, 2004.

[14] W. S. Jang and M. J. Skibniewski. "Wireless Network-based Tracking and Monitoring on Project Sites of Construction Materials", The 9th International Conference on Modern Building Materials, Structures and Techniques, 2007.

[15] S. Khan and M. Shah, "A Multiview Approach to Tracking People in Crowded Scenes Using a Planar Homography Constraint", Computer Vision ECCV 2006, vol. 3954, pp. 133-146, 2006.

[16] M. Kushwaha, O. Songhwai, I. Amundson et al., "Target Tracking in Heterogeneous Sensor Networks Using Audio and Video Sensor Fusion," Multisensor Fusion and Integration for Intelligent Systems, 2008.

[17] Y. Lao, J. Zhu and Y. Zheng, "Sequential Particle Generation for Visual Tracking", Circuits and Systems for Video Technology, IEEE Transactions on , vol. 19, no. 9, pp. 1365-1378, Sept., 2009.

[18] G. Mao, B. Fidan and B. Anderson, "Wireless Sensor Network Localization Techniques", Computer Networks, vol. 51, no. 10, pp. 2529-2553, 2007.

[19] L. M. Ni, Z. Dian and M. R. Souryal, "RFID-based Localization and Tracking Technologies", IEEE Wireless Communications, vol. 18, no. 2, pp. 45-51, April 2011.

[20] A. H. Sayed, A. Tarighat and N. Khajehnouri, "Network-based Wireless Location: Challenges Faced in Developing Techniques for Accurate Wireless Location Information", IEEE Signal Processing Magazine, vol. 22, no. 4, pp. 24-40, July 2005.

[21] S. Sun and Z. Deng, "Multi-sensor Optimal Information Fusion Kalman Filter", Automatica, vol. 40, issue 6, pp. 1017-1023, June 2004.

[22] A. Teymorian, W. Cheng, L. Ma, X. Cheng, X. Lu and Z. Lu, "3D Underwater Sensor Network Localization", in IEEE Transactions on Mobile Computing, pp. 1610-1621, Vol. 8, No. 12, December 2009.

[23] The Independent, "Bombers used unregistered mobiles to stay hidden", http://www.independent.co.uk/news/uk/home-news/bombers-used-unregistered-mobiles-to-stay-hidden-2106616.html, 2010.

[24] J. Wang, R. Ghosh and S. Das, "A Survey on Sensor Localization", Journal of Control Theory and Applications, vol. 8, issue 1, pp. 2-11, 2010.

[25] H. Wang, C. Tan and Q. Li, "Snoogle: A Search Engine for Physical World", IEEE INFOCOM, 2008.

[26] H. Wu, M. Siegel, R. Stiefelhagen and J. Yang, "Sensor Fusion Using Dempster-Shafer Theory", IEEE Instrumentation and Measurement Technology Conference, 2002.

[27] A. Yilmaz, O. Javed and M. Shah, "Object Tracking: A Survey", ACM Comput. Surv. vol. 38, no. 4, article 13, December 2006.

[28] Z. Yu, J. Teng, X. Bai, D. Xuan and W. Jia, "Connected Coverage in Wireless Networks with Directional Antennas", IEEE INFOCOM, 2011.

[29] L. A. Zadeh, "Fuzzy Sets", Information and Control, vol. 8, no. 3, pp. 338-353, 1965.

[30] J. Zhu, J. Teng, D. Xuan and Y. Zheng, "Effective Visual Tracking with Electronic Localization by Directional Antennas", IEEE NAECON, 2011.