A Novel Sequence Representation for Unsupervised Analysis of Human Activities

Raffay Hamid, Siddhartha Maddi, Amos Johnson, Aaron Bobick, Irfan Essa, Charles Isbell

College of Computing, Georgia Institute of Technology - Atlanta, GA, USA
{*raffay, maddis, amos, afb, irfan, isbell*}*@cc.gatech.edu*

Abstract

We present a novel activity representation as bags of event n-grams to extract global structural information of activities using their local event statistics. Exploiting this representation, we present a computational framework for unsupervised activity-class discovery, activity classification and anomalous activity detection. To this end, we model activity-classes as maximally similar activity-cliques in an edge-weighted graph of activities, and present a graph-theoretic method for their efficient discovery. Moreover, to detect irregular behaviors in active environments, we formulate a definition of anomalous activities, and propose an information theoretic method to explain the detected anomalies in a maximally informative manner. Finally, for the purposes of online activity classification and anomaly detection, we model the discovered activity-classes as variable memory Markov chains, and propose a method to find their constituent subsequences that are maximally exclusive amongst the discovered activity-classes. Results over data-sets collected from multiple environments are presented to demonstrate the competence of our framework.

Key words: Temporal Reasoning; Scene Analysis; Computer Vision.

1 Introduction

Consider a loading dock where activities such as the delivery or pick-up of packages take place. Analysis of what is happening in such active environments remains an important question, that will impact the development of systems for automatic surveillance and scene understanding. Our goal in this work is to introduce a novel representation that facilitates the extraction of activity-structure with minimal supervision. Exploiting this representation, we propose a computational framework to discover the various activity-classes taking place in an environment, to classify new activity instances as members

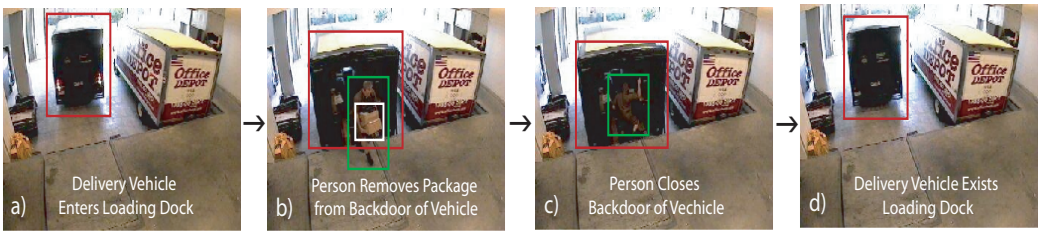


Fig. 1. **Key Frames of Example Events -** The figure shows an example delivery activity in a loading dock environment. The key-objects whose interactions define these events are shown in different colored blocks.

of one of the discovered activity-classes, and to automatically detect if the new activity instance is anomalous with respect to the general characteristics of its membership class.

Activities can be interpreted at different levels of abstraction [7]. Consider for example the activity of walking. One way of analyzing this activity is to focus directly on the motion-properties of image pixels [40]. Void of any semantic connotations of walking, this perspective strictly considers the pixel-statistics of local spatio-temporal regions of an image sequence. In sharp contrast is the activity-view which analyzes activities in terms of qualitative physical descriptions. With this perspective, the activity of walking can be described as putting one foot in front of another while keeping one foot on the ground at all time [34]. The granularity at which one chooses to decompose an activity offers a tradeoff between the prior knowledge needed to describe it, and the expressiveness of the activity-descriptors. This tradeoff is dictated by the dynamics of the activity, and the environment in which it is being performed. The activity dynamics may be simple or complex depending on the objects that are involved, and their temporal characteristics. Similarly, the environment in which the activity is being performed may be more or less constrained based on the variability of its physical properties.

To develop a perspective on activity decomposition, we begin with the assumption that an active environment consists of various **key-objects**, that provide the functionalities required for the execution of different activities in that environment. In the aforementioned example-scenario of a loading-dock, these key-objects may include delivery-vehicles, loading-carts, entrance-doors *etc.* For this work, we assume that a list of such key-objects for the environment under consideration is given *a priori*. We define an **event** as a particular interaction amongst a subset of these key-objects over a certain duration of time. We define the set of events that can be performed in an environment as the **event vocabulary**, and assume that this vocabulary is known *a priori*. Furthermore, we define an **activity** as a finite sequence of discrete events, and the temporal order of these events as **activity structure**. Key frames of events of a delivery activity in a loading dock are shown in Figure 1.

A majority of the previous work done in modeling human activities has focused on situations where the structure of the activity being modeled is known *a priori*. Context Free grammars [27], Expectations Grammars [36] and various derivates of probabilistic graphical models [25] [24] [14] *etc.* are examples of such fundamentally grammar based approaches. However, for many unconstrained settings, the activity structure is generally not known *a priori* [13]. Therefore there is a need for representations that can be used to learn this structure with minimal supervision.

To this end, we propose to represent activity sequences as histograms of their event n-grams where an n-gram is a contiguous activity-subsequence of length n. The intuition behind using such a representation is that it is sufficient to capture the global structural signature of an activity by looking at its local event subsequences. This intuition is in similar flavor with that of Allen's Algebra, which proposes that a small number of temporal relations are sufficient to encode the structure of temporal sequences [2]. Our work is an attempt to learn such temporal relations in an unsupervised manner. This idea has been previously used to capture structural information for object [18], speech [41], and document analysis [45].

We assume that activities taking place in an unconstrained environment do not span the activity space uniformly. Rather, there exist disjunctive activity-classes, that are essential for the analysis of activities. This assumption is consistent with how humans, in the face of a new piece of information, first classify it into an existing set of classes [44], and then compare it to the previous class-members to analyze how it varies in relation to the general characteristics of the membership class [46]. Adopting this perspective for unsupervised activity analysis calls for discovering the various activity-classes taking place in an active setting. There exists an inherent tradeoff between the cohesiveness and cardinality of these discovered classes [30]. One way of achieving a balance between these two opposing factors is to model a corpus of activity instances as a completely connected edge weighted graph of activities, and model activity-classes as maximally similar activity cliques in this graph. Here we propose an efficient method to discover such activity-cliques.

Activity-class discovery and characterization are precursors to anomaly detection [11] [17]. Most of the previous approaches for vision-based anomaly detection have focused on model-driven anomaly recognition [25] [10][32]. For any reasonably unconstrained situation however, anomalies are hard to define *a priori* [47]. Here, we first learn an instance-based model [1] of regular activities, and then detect an anomaly based on its dissimilarity from regular activities. Furthermore, we propose an information-theoretic method to find maximally informative features that distinguish an anomaly from the regular members of its membership class. Such explanations can be useful for large-scale surveillance systems. A block-diagram illustrating the general overview

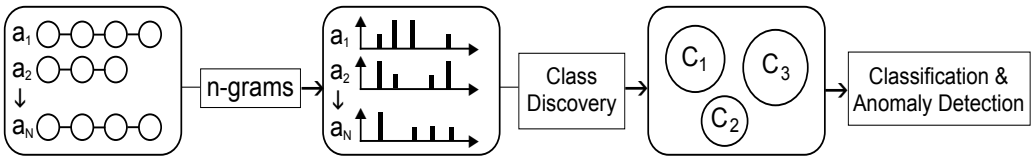


Fig. 2. **General Framework -** Starting with activities as finite sequences of discrete events, we (**1**) extract contiguous event subsequences of length n, *i.e.* event n-grams. Considering a corpus of activities as an edge-weighted activity-graph, we (**2**) discover activity-classes as maximal cliques in this graph. Given a new activity instance, we (**3**) classify it to one of the discovered classes, and find its distinguishing features.

of our proposed framework is given in Figure 2.

For the example scenarios that we have considered in this work, the start and end of each activity instance are known explicitly. There are however quite a few problem domains where such demarcations are not so clear. For such situations, it is crucial to find concise characterizations of the discovered activity-classes that could be used for online activity classification and detection of various anomalies. We view this problem as that of finding recurrent event subsequences that are maximally mutually exclusive amongst the various activity-classes, and can be used to find different irregularities. We call such recurrent event subsequence as *event motifs* (formally defined in § 7.1) and use variable-memory Markov chains for their discovery [43].

This paper undertakes a detailed analysis of some of our preliminary work in [20] and [21]. We start in Section 2 by reviewing the previous work related to the problem at hand, pointing out how our approach is different from the previously proposed methods. We explain in Section 3 our proposed activity representation of event n-grams, and present an empirical analysis of their discriminative power and sensitivity to sensor noise as a function of class overlap. To this end, we propose a novel method to systematically control class-overlap while capturing event dependence in activity sequences over variable temporal durations. Crucial to analyze the competence of sequence representations, our proposed method offers applications to the more general problem of sequence analysis. Exploiting event n-grams, in Section 4 we show how the notion of maximal cliques in edge-weighted activity-graphs can be used to efficiently discover activity-classes in an unsupervised manner. In Section 5, we explain how these discovered activity-classes can be used for activity classification, anomalous activity detection as well as their explanation. Section 6 explains the experimental results for our proposed framework. The characterization of the discovered activity classes for the purposes of online activity classification and anomaly detection is presented in Section 7. Section 8 explains the results for our proposed framework for event motif discovery. The conclusions and future directions of this work are explained in Section 9.

2 Related Work

The problem of everyday human activity analysis has been studied in various contexts, including computational perception [8], ubiquitous computing [15], and robotics [48]. Much has been written about activity decomposition and the role of knowledge in the perception of motion [7], where scientists have worked on understanding the psychological [49] as well as computational basis of how motion is perceived. [52] [50].

One of the key problems in this regard is finding representations that are robust and efficiently computable. Most of the previous approaches towards this end have been fundamentally grammar-driven (see *e.g.* [27] [9] [33] and the references therein). Such representations explicitly model the activity-structure followed by the learning of model parameters given some training data. For reasonably unconstrained settings however, the activity structure is generally not known *a priori*. Therefore, there is a need for representations that encode this structure with minimal supervision. In this work we propose to treat activities as bags of event n-grams, and attempt to extract their global structural information, by analyzing their local event statistics.

While discovering activity-classes has been previously explored in such fields as network intrusion detection [32], it has only recently gained attention in vision-based activity analysis [35]. Our approach towards this problem is novel in a few key aspects. Unlike [25] which require *a priori* expert knowledge to model the activity-classes in an environment, we propose to discover this information in an unsupervised manner. Since event-monograms, as used in [53] and [47], do not capture the temporal information of an activity, we propose to use higher order event n-grams.

Moreover, unlike previous approaches, our framework models activity classes as edge-weighted maximal cliques in a completely connected graph of some given activity-instances. Finding maximal cliques in edge-weighted graphs is a classic graph theoretic problem [4] [42]. Here, we adopt the recently proposed approximate approach of iteratively finding dominant sets of maximally similar nodes in a graph (equivalent to finding maximal cliques) [39]. Besides providing an efficient approximation to finding maximal cliques, the framework of dominant sets naturally provides a principled measure of the cohesiveness of a class as well as a measure of node participation in its membership class.

Most of the previous attempts to tackle the problem of anomaly detection have focused on model-based anomaly recognition [25] [26]. For reasonably unconstrained situations however, anomalies are hard to completely define *a priori*. Rather than modeling anomalies themselves, in this work we propose to model the regular activity classes and detect anomalous activities based on

their distance from learned models of regular behaviors in the environment.

A concise characterization of these discovered activity-classes is imperative, both from a representational as well as a discriminative perspective. This is particularly important in situations where the start and end of different activities is not explicitly marked, and there is a need to perform online classification and anomaly detection. While previously proposed instance-based approaches in this regard [29] [47] focus on the representational aspects of the problem, they are not necessarily discriminative. Moreover, these approaches only consider activities at a very global scale, not incorporating the more local information. To this end, we formalize this problem as finding predictably recurrent event motifs using variable memory Markov chains.

Numerous solutions to the problem of discovering important recurrent motifs in the fields of Bioinformatics and String Analysis have been previously proposed (see *e.g.* [37] [6] [12] and the references therein). Work done in [51] and [43] present techniques for learning variable memory Markov chains from training data in an unsupervised manner. Here, we extend the work done in [51] to handle data from multiple classes, finding motifs that are maximally mutually exclusive amongst activity-classes. Instead of sequentially finding individual subsequences and masking them out from the sequences as proposed in [5], our scheme simultaneously finds all the subsequences in the data in one pass, allowing to find partially overlapping subsequences.

3 Representation - Activities as Bags of Event n-grams

As models of activity structure for unconstrained environments are generally not available *a priori* [13], representations that can encode this structure with minimal supervision are needed. Looking at an activity as a sequence of discrete events, two important quantities emerge:

- *Content* - events that span the activity, and
- *Order* - the arrangement of the set of events.

This treatment of an activity is similar to the representation of a document known as the Vector Space Model (VSM) [45], where a document is represented as a vector of its word-counts. While approaches such as VSM capture the content of a sequence in an efficient way, they completely ignore its order. Because the word content in documents often implies causal structure, this is usually not a significant challenge. Activities however are not fully defined by their event-content alone; rather, there are preferred or typical event-orderings [37]. Therefore a model to capture event order in a more explicit manner is needed.

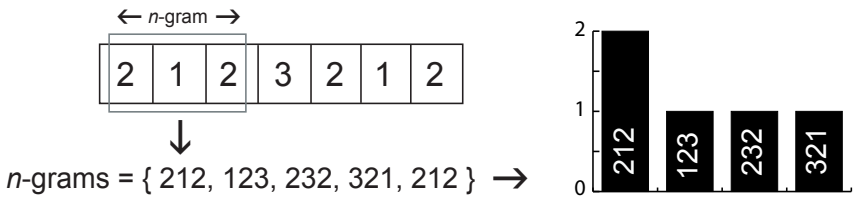


Fig. 3. **Illustration of n-grams -** Transformation of an example activity from sequence of discrete events to histogram of event n-grams. Here the value of n is shown to be equal to 3.

To this end, we consider activities as histograms of event n-grams where an n-gram is a contiguous subsequence of activity of length n. By sliding a window of length n over an activity, we can find all the event n-grams contained in that activity. We can then represent the activity as counts of these extracted n-grams. For the example shown in Figure 3, the value of n is set equal to 3. For the purposes of this work, we assume that the start and end of each activity are known, and that every activity is finished before another is started. In cases where multiple agents are simultaneously performing different events, the mapping between the different events to the different agents is assumed to be known.

It is evident that higher values of n would capture the temporal order information of events more rigidly, and would entail a more discriminative representation. However, as n increases, the dimensionality of the space grows exponentially. This highlights the importance of selecting a reasonable value of n which would capture event dependence in an environment to a sufficient extent, while spanning a minimal dimensional space. Finding the optimal value of n is a non-trivial problem, and while there exist previously proposed methods in sequence analysis to learn the optimal value of n [22] [31], this particular problem is beyond the scope of this work. For our purposes, we fixed n to a constant value of 3.

3.1 Empirical Analyses of n-grams using Simulation Data

Representations such as n-grams can be thought of as a means to extract different sequential features from an activity sequence. It is essential to analyze how well can such a feature space discern between members of different classes with respect to some ground-truth notion of class-overlap. Moreover, since for any sensor-based perceptual system, the observations are always prone to sensor-noise, the efficacy of a representation is a function of how sensitive it is to such sensor-noise. With this perspective at hand, we now present empirical analyses of n-grams in terms of their discriminative power and noise sensitivity as a function of class-disjunction and noise perturbation. The analyses presented here are based on simulated data, the details for which follow.

Events in human activities depend on preceding events over variable durations [38]. To simulate this variable length event dependence, we model activity classes as variable memory Markov chains (*VMMC*) [51]. One way of encoding such a *VMMC* is by using a probabilistic tree [19], where each node represents any one of the members of the event vocabulary, while each edge represents the probability of traversing to its child from its parent. The topology of a tree encodes the variable temporal dependence between different events. Given two identical trees, the sequences generated from them would have same statistical properties. However, as we increasingly perturb their edge probabilities, the resulting sequences generated would have increasingly different event statistics. Using this behavior to model the disjunction between the sequences of a couple of activity-classes, we first outline a novel algorithm regarding how to systematically control this class disjunction.

3.1.1 Systematic Control Over Class Disjunction

We begin by constructing a complete tree T with depth equal to d. Randomly selecting half of the leaf-nodes of T, we iteratively attach them to its remaining half. The *VMMC* for class-1 is completed by assigning edge-probabilities of T by sampling from a normal distribution with zero mean and unit variance ($\mathcal{N}(0,1)$). *VMMC* for class-2 is constructed by first forming an exact copy of *VMMC* of class-1, followed by perturbing edge probabilities of top $\eta\%$ edge-paths of *VMMC* for class-1. The algorithm is outlined in Algorithm 1, and figuratively illustrated in Figure 4.

3.1.2 Simulation Data:

For a symbol vocabulary $||\Sigma|| = 5$ and modal depth equal to 3, we generated 10 different topologies of *VMMCs*. For each topology, we generated sequences for 2 classes with percent overlap decreasing from complete overlap to complete non-overlap with increments of 10%. For each of these 100 trials, we generated

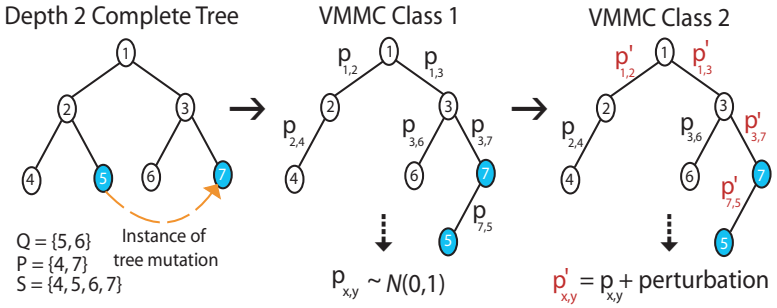


Fig. 4. **Illustration of Algorithm 1** - We begin by constructing a complete tree of depth d. $\mathcal{P}$ and $\mathcal{Q}$ are selected from leaf-set $\mathcal{S}$. Probabilities of VMMC-1 are sampled from $\mathcal{N}(0,1)$. VMMC-2 is constructed by perturbing probabilities of VMMC-1.

Algorithm 1 Construct *VMMC's* $\mathcal{V}_1$ and $\mathcal{V}_2$

Require: Symbol vocabulary k, modal depth d, number of topological operations I, and % node perturbation η

Construct $\mathcal{V}_1$ as complete tree of depth d with leaf-set $\mathcal{S}$
Randomly construct $\mathcal{P} \subseteq \mathcal{S}$ where $||\mathcal{P}|| = ||\mathcal{S}||/2$
Construct $\mathcal{Q} \equiv \mathcal{S} \setminus \mathcal{P}$
for $i = 1$ to I **do**
    Sample a member of $\mathcal{Q}$. Detach it from its parent. Attach it to a randomly selected member of $\mathcal{Q}$.
end for
Sample edge probability of $\mathcal{V}_1$ from $\mathcal{N}(\mu, 1)$ distribution

Construct $\mathcal{V}_2$ as an exact copy of $\mathcal{V}_1$
Sample edge probability of η% nodes of $\mathcal{V}_2$ from $\mathcal{N}(\mu, 1)$

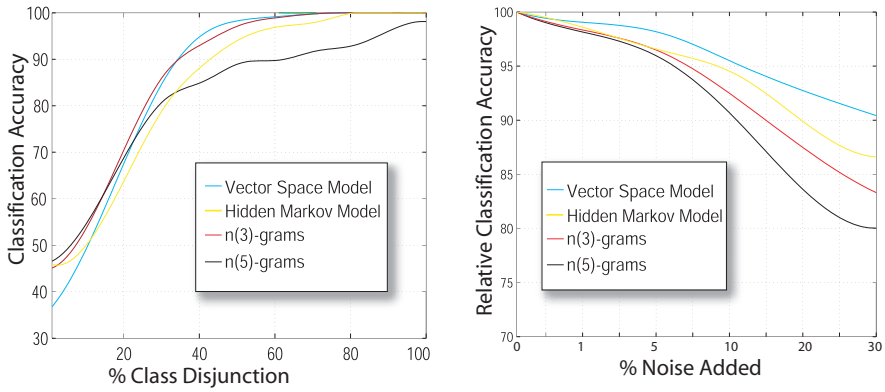
75 sequences each of length 100, randomly selecting two-thirds for the training data and the rest for testing.

3.1.3 Discriminability Analysis

For data generated as described in § 3.1.2, and using similarity metric defined later (Equation 1), the nearest neighbor classification results are given in Figure 5-**a**. It is evident that for substantive class overlap, higher values of n seem to capture activity structure more rigidly, entailing a more discriminative representation. However, since accurate density estimation for higher value n-grams require exponentially greater amount of data, Vector Space Model seems to outperform 3- and 5-grams in cases where the 2 classes are more disjunctive.

3.1.4 Noise Sensitivity Analysis

We now analyze noise sensitivity of n-grams as a function of noise added as *Insertion*, *Deletion*, *Transposition* and *Substitution* of symbols. For data generated as described in § 3.1.2, we cumulatively added all four types of noises with a uniform prior on each, and noise likelihood ranging monotonically from 0 to 30%. Using noisy data, the classification results for different representations relative to their noise free performance is given in Figure 5-**b**. It is evident that representations that capture event order information more rigidly, are more sensitive to sensor noise. This underlines an inherent tradeoff between the ability of a representation to explicitly capture sequence-structure, and its robustness to sensor noise. Considering both these opposing factors, it seems that tri-grams ($n = 3$) provide a reasonable balance between the two opposing factors. This is particularly true for relatively small class-overlap.



(a) (b)

Fig. 5. **a - Discriminative Prowess** - Classification accuracy as a function of class-overlap. **b - Noise Sensitivity**- Classification for various representations relative to their noise free performance.

4 Unsupervised Activity-Class Discovery

In this section, we use our proposed activity-representation of event n-grams to discover the various activity classes taking place in an environment. As these activity-classes have high internal similarity and low external similarity, we first need to define our notion or activity similarity before we formalize a method of discovery for these activity-classes.

4.1 Activity Similarity Metric

Sequence comparison is a well-studied problem with numerous applications [19]. Our view of the similarity between a pair of activity sequences consists of two factors, the core structural differences and differences based on the frequency of occurrence of event n-grams.

The core structural differences relate to the distinct n-grams that occurred in either one of the sequences in a sequence-pair, but not in both. For such differences, the number of mutually exclusive n-grams is of fundamental interest. On the other hand, if a particular n-gram is present in both the sequences, the only discrimination that can be drawn between the sequence-pair is purely based on the frequency of the occurrence of that n-gram.

This intuition can be formalized as follows. Let A and B denote two sequences of events, and let their corresponding histogram of n-grams be denoted by H_A and H_B. Let Y and Z be the sets of indices of n-grams with counts greater than zero in H_A and H_B respectively. Let α_i denote different n-grams. $f(\alpha_i|H_A)$ and

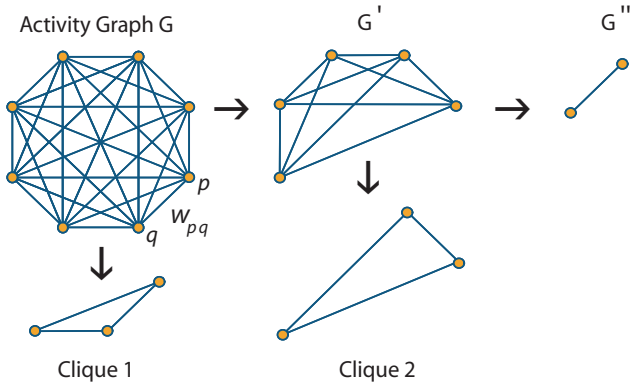


Fig. 6. **Illustration of Activity Class Discovery -** Activity-instances are represented as a completed connected, edge-weighted activity graphs G. The edge-weight $w_{p,q}$ between nodes p and q is computed using Equation 1. Maximal cliques of activity-nodes are iteratively found and removed from the activity-graph, until there remain no non-trivial maximal cliques. These maximal cliques correspond to activity-classes comprising of mutually similar activity instances.

$f(\alpha_i|H_B)$ denote the counts of α_i in sequences A and B respectively. We define similarity between two activities as:

$$\text{sim(A,B)} = 1 - \kappa \sum_{i \in Y,Z} \frac{|f(\alpha_i|H_A) \ - \ f(\alpha_i|H_B)|}{f(\alpha_i|H_A) \ + \ f(\alpha_i|H_B)} \tag{1}$$

where $\kappa = 1/(||Y|| + ||Z||)$ is the normalizing factor, and $||\cdot||$ computes the cardinality of a set. While our proposed similarity metric conforms to: (1) the property of Identity of indiscernibles, (2) is commutative, and (3) is positive semi-definite, it does not however follow the triangular inequality, making it a divergence rather than a true distance metric.

4.2 Activity-Class Discovery

It is argued that while facing a new piece of information, humans first classify it into an existing class [44], and then compare it to the previous class members to understand how it varies in relation to the general characteristics of the membership class [46]. Using this perspective as our motivation, we represent an activity space by a set of mutually disjunctive classes, and then detect a new activity as a regular or an anomalous member of its membership class.

4.2.1 Activity-Class as Maximal Clique

Given K activity-instances, we consider this activity-set as an undirected edge-weighted graph with K nodes, each representing a histogram of n-grams of one of the K activity-instances. The weight of an edge is the similarity between a

pair of nodes as defined in Equation 1. We can now formalize the problem of discovering activity-classes as searching for edge-weighted maximal cliques [1] in the graph of K activity-instances [4]. We begin by finding the first maximal clique in the activity-graph, followed by removing that set of nodes from the graph, and iteratively repeating this process with the remaining set of nodes, until there remain no maximal cliques in the graph. The leftover nodes after the removal of maximal cliques are dissimilar from most of the regular nodes, and are considered as being anomalous (see Figure 6 for illustration).

4.2.2 Maximal Cliques using Dominant Sets

As combinatorially searching for maximal cliques in an edge-weighted undirected graph is computationally hard, numerous approximations to the solution of this problem have been proposed [42]. For our purposes, we adopt the approximate approach of iteratively finding *dominant sets* of maximally similar nodes in a graph (equivalent to finding maximal cliques) as proposed in [39]. Besides providing an efficient approximation to finding maximal cliques, the framework of dominant sets provides a principled measure of cohesiveness of a class as well as a measure of node participation.

Let the data to be clustered be represented by an undirected edge-weighted graph with no self-loops $G = (V, E, \vartheta)$ where V is the vertex set $V = \{1, 2, ...K\}$, $E \subseteq V \times V$ is the edge set, and $\vartheta : E \rightarrow \mathbb{R}^+$ is the positive weight function. The weight on the edges of the graph are represented by a corresponding $K \times K$ symmetric similarity matrix $A = (a_{ij})$ defined as:

$$a_{ij} = \begin{cases} \text{sim}(i,j) \text{ if } (i,j) \in \text{ E} \\ 0 \qquad\qquad\quad \text{otherwise} \end{cases} \tag{2}$$

Here $\text{sim}(i, j)$ is computed using our proposed notion of similarity as defined in Equation 1. To quantize the cohesiveness of a node in a cluster, we define its "average weighted degree". Let $S \subseteq V$ be a non-empty subset of vertices and $i \in S$, such that,

$$\text{awdeg}_{\text{S}}(i) = \frac{1}{||\text{S}||} \sum_{j \in \text{S}} a_{ij} \tag{3}$$

and

$$\Phi_{\text{S}}(i,j) = a_{ij} - \text{awdeg}_{\text{S}}(i) \quad \text{for } j \notin \text{S} \tag{4}$$

Intuitively, $\Phi_{\text{S}}(i, j)$ measures the similarity between nodes j and i, with respect to the average similarity between node i and its neighbors in S. Note that

[1] Recall that a subset of nodes is a *clique* if all its nodes are mutually adjacent; a *maximal* clique is not contained in any larger clique; a *maximum* clique has largest cardinality.

$\Phi_{\mathrm{S}}(i,j)$ can either be positive or negative.

We now consider how weights are assigned to individual nodes. Let $S \subseteq V$ be a non-empty subset of vertices and $i \in S$. The weight of i with respect to S is given as:

$$w_{\mathrm{S}}(i) = \begin{cases} 1 & \text{if } ||\mathrm{S}|| \ = \ 1 \\ \sum\limits_{j \in \mathrm{S}\setminus\{i\}} \Phi_{\mathrm{S}\setminus\{i\}}(j,i) w_{\mathrm{S}\setminus\{i\}}(j) & \text{otherwise} \end{cases} \tag{5}$$

Intuitively, $w_S(i)$ gives a measure of the overall similarity between vertex i and the vertices of $S\backslash\{i\}$ with respect to the overall similarity among the vertices in $S\backslash\{i\}$. We are now in a position to define *dominant sets*. A non-empty sub-set of vertices $S \subseteq V$ such that $W(T) > 0$ for any non-empty $T \subseteq S$, is said to be *dominant* if:

- $w_{\mathrm{S}}(i) > 0$, $\forall\ i \in \mathrm{S}$, *i.e.* internal homogeneity
- $w_{\mathrm{S}\cup\{i\}}(i) < 0\ \forall\ i \notin \mathrm{S}$, *i.e.* external inhomogeneity.

Effectively, we can state that the dominant set in an edge-weighted graph is equivalent to a cohesive cluster of vertices in that graph. Because solving Equation 5 combinatorially is infeasible, we use a continuous optimization technique called replicator dynamics (for details, see [39]).

5 Activity Classification and Anomaly Explanation

Given $||C||$ discovered activity-classes, we are interested in finding if a new activity instance is regular or anomalous. Each member j of an activity-class c has some weight $w_c(j)$, that indicates the participation of j in c. We compute the similarity between a new activity-instance τ and the previous members of each class by defining a function $A_c(\tau)$ as:

$$A_c(\tau) = \sum_j sim(\tau, j) w_c(j) \quad \forall j \in c \tag{6}$$

Here $w_c(j)$ is the same as defined in Equation 5. A_c represents the average weighted similarity between the new activity-instance τ and any one of the discovered classes c. The selected membership class c^* is found as

$$c^* = \arg\max_{\forall c} A_c(\tau) \tag{7}$$

Once the membership decision of a new test activity has been made, we now focus our attention on deciding whether the new class member is regular or anomalous. Intuitively speaking, we want to decide the normality of a new instance based on its closeness to the previous members of its membership activity-class. This is done with respect to the average closeness between all the previous members of its membership class. Let the function $\Gamma(\tau)$ be:

$$\Gamma(\tau) = \sum_{j \in c^*} \Phi_{c^*}(j, \tau) w_{c^*}(j) \tag{8}$$

where Φ is defined by Equation 4. We define a new class member τ as regular if $\Gamma(\tau)$ is greater than a particular threshold. The threshold on $\Gamma(\tau)$ is learned by mapping all the anomalous activity instances detected in the training activity-set to their closest activity-class (using Equation 6 & 7), and computing the value of Γ for both regular and anomalous activity instances. We can now observe the variation in false acceptance rate and true positives as a function of the threshold Γ. This gives a "Receiver Operating Curve" (ROC). The area under ROC is indicative of the confidence in our detection metric $\Gamma(\tau)$ [28]. Based on our tolerance for true and false positive rates, we can choose an appropriate threshold.

5.1 Anomaly Explanation

Explanation of the detected anomalous activities is a function of characterization of the general properties of an activity class. One way of characterizing these properties is to find the best representative or typical member of a class [29]. The question of typicality is closely related to the similarity of a node to other members of a class. The problem has been previously approached as finding the node with min-max distance from other nodes [16], or the node with maximum in-degree [23]. Such approaches however either assume the clusters to be well behaved, or take a very local view of a node's similarity to its neighbors.

5.2 Activity Class Modeling

Following [29], we propose the idea of typical nodes (mentioned as "authoritative sources" in [29]) and "similar to typical (STT)" nodes (mentioned as "hubs" in [29]). Typical and STT nodes exhibit a mutually reinforcing relationship - a good STT node is one which is closer to a Typical node, while a Typical node is one closer to more STT nodes. Following [29], we associate a non-negative Typicality weight x^p and a non-negative STT weight y^p to each node in the cluster where p denotes the index of nodes in a cluster. Naturally,

if p is closer to many nodes with large x values, it should receive a large y value. On the other hand if p is closer to nodes with large y values, it should receive large x value. We define two coupled processes to update weights x^p and y^p iteratively, *i.e.*

$$x^p \leftarrow \sum_{q:(q,p)\in E} y^p \quad \text{and} \quad y^p \leftarrow \sum_{q:(q,p)\in E} x^p \tag{9}$$

As we iterate the above two equations k times in the limit $k \leftarrow \infty$, x^p and y^p converge to x^* and y^*. The node which has the largest component in the converged vector x^* would correspond to the node which has the greatest Typical weight and hence is the best representative of the nodes of clusters. x^* can be computed from the Eigen Analysis of the matrix A^TA where A is the symmetric similarity matrix of all the nodes of the cluster. Essentially x^* is the principal eigenvector (the one with greatest corresponding Eigen value) of A^TA, the largest component of which corresponds to the Typical Node of the cluster (for the proof, please refer to [29]).

5.2.1 Explanatory Features

For large scale surveillance systems, it is imperative to find the features that can be used to explain an anomalous activity in a maximally-informative manner. We are interested in features of an activity-class that have minimum entropy, and occur frequently. The entropy of a tri-gram indicates the variation in its observed frequency, which in turn indicates the confidence in the prediction of its frequency. The frequency of occurrence of a tri-gram suggests its participation in an activity-class. We want to analyze the extraneous and the pertinent features in an activity sequence that made it anomalous with respect to the most explanatory features of the regular members of the membership activity-class. We now construct our approach mathematically (a figurative illustration is given in Figure 7).

Let α_i denote a particular tri-gram i for an activity, and c denote any of the $||C||$ discovered activity-classes. If R denotes the typical member of c as described in §5.2, and τ denotes a new activity-class member detected as being anomalous, then we can define the difference between their counts for α_i as:

$$D(\alpha_i) = f_R(\alpha_i) - f_\tau(\alpha_i) \tag{10}$$

where $f(\alpha_i)$ denotes the count of a tri-gram α_i. Let us define the distribution of the probability of occurrence of α_i in c as:

$$P_c(\alpha_i) = \frac{\sum\limits_{k\in c} f_k(\alpha_i)}{\sum\limits_{i=1}^{M}\sum\limits_{k\in c} f_k(\alpha_i)} \tag{11}$$

where M represents all the non-zero tri-grams in all the members of activity-class c. Let us define multi-set χ_c^i as:

$$\chi_c^i = \{f_k(\alpha_i) | k \in c\} \tag{12}$$

We can now define probability $Q(x)$ of occurrence of a particular member $x \in \chi_c^i$ for α_i in c as:

$$Q(x) = \psi \sum_{j \in c} \begin{cases} 1 \ if \ f(\alpha_i) = \ x \\ 0 \qquad \text{otherwise} \end{cases} \tag{13}$$

where ψ is the normalization factor. Let us define Shannon's Entropy of a tri-gram i for an activity-class c by $H_c(\alpha_i)$ as:

$$H_c(\alpha_i) = \sum_{x \in \chi_c^i} Q_c(x) ln(Q_c(x)) \tag{14}$$

We can now define the notion of *predictability*, $\text{PRD}_c(\alpha_i)$, of the values of tri-gram α_i of cluster c as:

$$\text{PRD}_c(\alpha_i) = 1 - \frac{H_c(\alpha_i)}{\sum\limits_{i=1}^{M} H_c(\alpha_i)} \tag{15}$$

It is evident from this definition, that α_i with high entropy $H_c(\alpha_i)$ would have high variability, and therefore would have low predictability.

We define the explainability of a tri-gram $\alpha_i \in c$ that was frequently and consistently present in the regular activity-class as:

$$\xi_c^P(\alpha_i) = \text{PRD}_c(\alpha_i) P_c(\alpha_i) \tag{16}$$

Intuitively, ξ_c^P indicates how much an α_i is instrumental in representing a activity-class c.

Similarly, we can define the explainability of $\alpha_i \in c$ in terms of how consistently was it absent in representing c.

$$\xi_c^A(\alpha_i) = \text{PRD}_c(\alpha_i)(P_c^{\max}(\alpha_i) - P_c(\alpha_i)) \tag{17}$$

where $P_c^{\max}(\alpha_i)$ is the maximum probability of occurrence of any α_i in c.

The first term in both Equation 16 and 17 indicates how consistent α_i is in its frequency over the different members of a class. The second term in Equation 16 and 17 dictates how representative and non-representative α_i is for c respectively.

Given an anomalous member of a activity-class, we can now find the features that were frequently and consistently present in the regular members of the

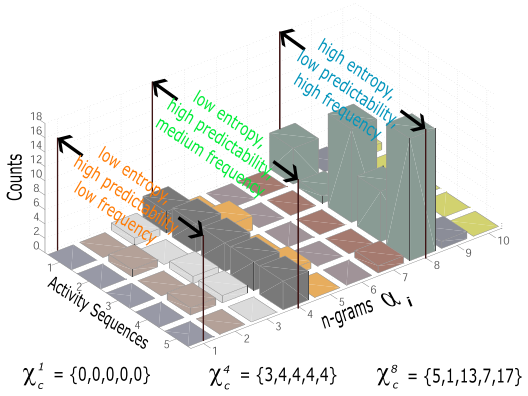


Fig. 7. Five simulated activity sequences are shown to illustrate the different concepts introduced in § 5.2.1. α_1 has low value of P_c, its entropy H_c is low and therefore its predictability is high. α_4 has medium P_c, its entropy H_c is also low and its predictability is high. Finally α_8 has high P_c, but its entropy H_c is high which makes its predictability low. α_1 could be useful in explaining the extraneous features in an anomalous activity, while α_4 could be useful in explaining the features that were deficient in an anomaly.

activity-class, but were deficient in the anomaly τ. To this end, we define the function Deficient(τ) as:

$$\text{Deficient}(\tau) = \arg\max_{\alpha_i}[\xi_c^P(\alpha_i)D_c(\alpha_i)] \tag{18}$$

Similarly, we can find the most explanatory features that were consistently absent in the regular members of the membership activity-class but were extraneous in the anomaly. We define the function Extraneous(τ) as:

$$\text{Extraneous}(\tau) = \arg\min_{\alpha_i}[\xi_c^A(\alpha_i)D_c(\alpha_i)] \tag{19}$$

We can explain anomalies based on these features in two ways. Firstly, we can consider features that were deficient from an anomaly but were frequently and consistently present in the regular members. Secondly, we can consider features that were extraneous in the anomaly but were consistently absent from the regular members of the activity-class.

6 Results: Class Discovery, Classification & Anomaly Explanation

To test the competence of our proposed framework, experiments on extensive data-sets collected from two active environments were performed. For both experimental setups, the value of n for the n-grams was chosen to be equal to 3 (tri-grams). This is done to follow the 2^{nd} order Markov assumption, *i.e.* an n-gram would encode the past, present and the future events.

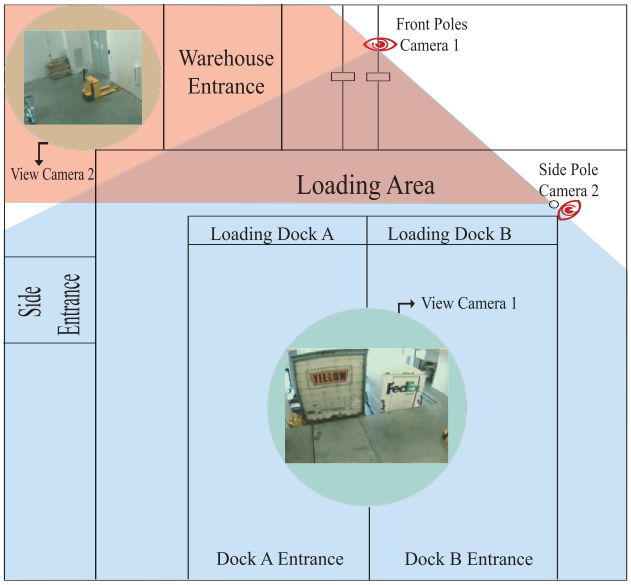


Fig. 8. A schematic diagram of the camera setup at the loading dock area with overlapping fields of view.

6.1 Loading Dock Environment - LDE

We collected video data at the Loading Dock area of a retail bookstore. To visually span the area of activities in the loading dock, we installed two cameras with partially overlapping fields of view. A schematic diagram with sample views from the two cameras is shown in Figure 8. Daily activities from 9a.m. to 5p.m., 5 days a week, for over one month were recorded. Based on our observations of the activities taking place in that environment, an event vocabulary of 61 events was constructed. Every activity has a known starting event, *i.e.* Delivery Vehicle Enters the Loading Dock and a known ending event, *i.e.* Delivery Vehicle Leaves the Loading Dock. We used 150 of the collected instances of activities, that were manually annotated using our defined event-vocabulary of 61 events. The 10 key objects whose various interactions constituted these 61 events were: Person, Cart, Delivery Vehicle(D.V.), Left Door of D.V., Right Door of D.V., Back Door of D.V., Package, Doorbell, Front Door of Building, Side Door of Building.

6.2 Discovered Activity Classes - Loading Dock Environment

Out of the 150 training activities, we found 7 classes (maximal cliques), with 106 activities as part of any one of the discovered class, while 44 activities being different enough to be not included into any non-trivial maximal clique. The visual representation for the similarity matrices of the original 150 activities and the re-arranged activities in 7 clusters is shown in Figure 9. Analysis of

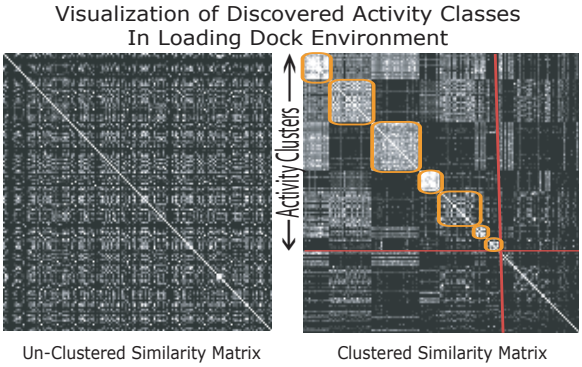


Fig. 9. Each row represents the similarity of a particular activity with the entire activity training set. White implies identical similarity while black represents complete dissimilarity. The activities ordered after the red cross line in the clustered similarity matrix were dissimilar enough from all other activities as to not be included in any non-trivial maximal clique.

the discovered classes reveals a strong structural similarity amongst the class members. A brief description of the discovered activity-classes is given in the following:

- **Class 1:** UPS® delivery-vehicles that picked up multiple packages using hand carts.
- **Class 2:** Pickup trucks and vans that dropped off a few packages without needing a hand cart.
- **Class 3:** Delivery trucks that dropped off multiple packages, with multiple people using hand-carts.
- **Class 4:** A mixture of car, van, and truck delivery vehicles that dropped off one or two packages without needing a hand cart.
- **Class 5:** Delivery-vehicles that picked up and dropped-off multiple packages using a motorized hand cart and multiple people.
- **Class 6:** Van delivery-vehicles that dropped off one or two packages without needing a hand cart.
- **Class 7:** Delivery trucks dropped off multiple packages using hand carts.

6.3 House Environment

To test our proposed algorithms on the activities in a house environment, we deployed 16 strain gages at different locations in a house, each with a unique identification code. These transducers register the time when the resident of the house walk over them. The data was collected daily for almost 5 months (151 days - each day being considered as an individual activity). Whenever the person passed near a transducer at a particular location, it was considered as the occurrence of a unique event. Thus our event vocabulary in this envi-

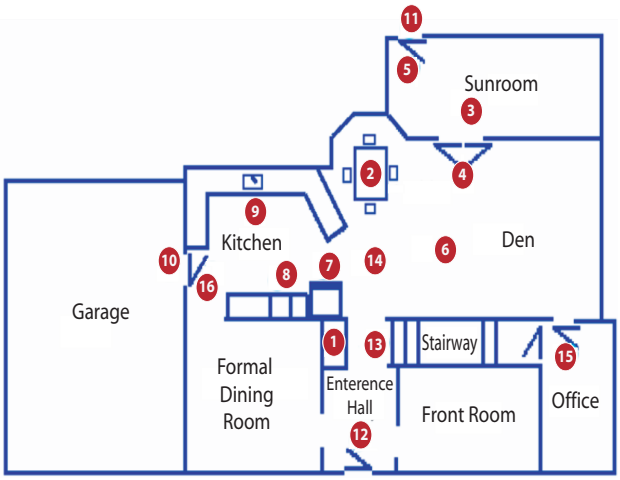


Fig. 10. A schematic diagram of the strain-gage setup in the house scenario. The red dots represents the positions of the strain gages.

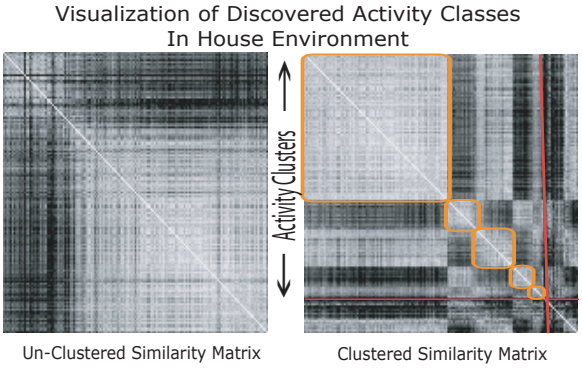


Fig. 11. Visualization of similarity matrices before and after class discovery for the House Environment.

ronment consists of 16 events. Figure 10 shows a schematic top-view of this environment.

6.4 Discovered Activity Classes - House Environment

Of the 151 activities captured over a little more than 5 months, we found 5 activity-classes (maximal cliques), with 131 activities as members of any one of the discovered class, and 20 activities being dissimilar enough not to be a part of any non-trivial maximal clique (see Figure 11). A brief description of the discovered activity-classes is given below:

- **Class 1:** Activities lasting for the entire length of days where the person's trajectory spans the entire house space. Most of the time was spent in the area around the Kitchen and the Dining Table.
- **Class 2:** The person moves from from kitchen to the stairway more often.

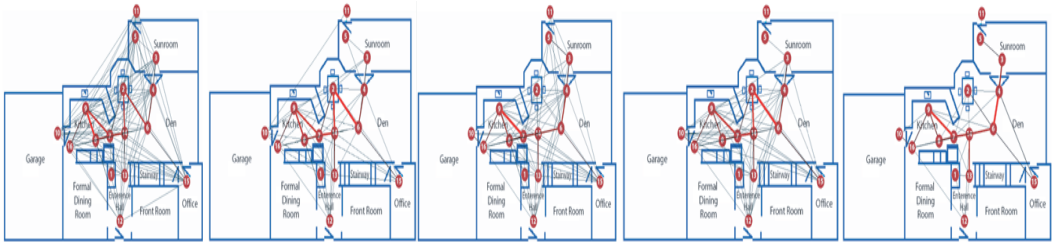


Fig. 12. Visualization of the structural differences between the discovered activity–classes. Thick lines with brighter shades of red indicate higher frequency.

Further more, as opposed to cluster 1, the person does not go from the Office to the Sum Room area.

- **Class 3:** The person spends more time in the areas of Den and the living-room. Moreover, he visits the Sun-room more often.
- **Class 4:** The person spends most of the day in Kitchen and Dining Room. The duration for which she stays in the house is smaller for this class.
- **Class 5:** The person moves from Dining Room to the Sun Room more often. The duration for which she stays in the house is significantly smaller than any other activity-class.

To illustrate the structural differences in the discovered activity-classes, a visualization of normalized frequency-counts of the person's trajectory between different locations is shown in figure 12.

6.5 Analysis of Discovered Classes

The discovered activity-classes both for the Loading Dock and the House data-sets, are semantically coherent and divide their respective activity space discriminatively. The fundamental differences between various classes in the Loading Dock environment are dictated by the fact whether the activities were of delivery or pick-up, how many people were involved in the activity, how many packages were moved, and what type of delivery vehicle was used. For the House environment, these differences pertain the time a person stays in house, and the time of the year it is.

Figures 9 and 11 show that the activities performed in the Loading Dock environment are structurally more well defined than those performed in the House environment. This is because our vocabulary for the Loading Dock environment consists of semantically meaningful events, which can encode the underlying activity structure efficiently. For the House environment, the events are simply the locations where a person went, and are not particularly designed to encode the underlying structure of the activities.

6.6 Detected Anomalies - Loading Dock Environment

We performed analysis on the anomalous activities exclusively for the Loading Dock scenario. Out of the 150 training activities, we found 7 classes (maximal cliques), with 106 activities as part of any one of the discovered class, while 44 activities being different enough to be not included into any non-trivial maximal clique. We now give a detailed explanation of how, using these initially detected anomalous activities, we can learn a threshold for detecting new anomalous activity-class members, how valid are these detected anomalies from a human view-point, and finally, what explanations did we get for detected anomalous activities based on the selected key-features of the activity-classes.

6.6.1 Analysis of Detected Anomalies

Analyzing the detected anomalous activities reveals that there are essentially two kinds of activities that are being detected as anomalous, (1) ones that are truly alarming, where someone must be notified, and (2) those that are simply unusual delivery activities with respect to the other regular activities. Key-frames for three of the truly alarming anomalous activities are shown in Figure 13. Figure 13-a shows a truck driving out without closing it's back door. Not shown in the key-frame is the sequence of events where a loading-dock personnel runs after the delivery vehicle to tell the driver of his mistake. Figure 13-b shows a delivery activity where a relatively excessive number of people unload the delivery vehicle. Usually only one or two people unload a delivery vehicle, however as can be seen from Figure 13-b, in this case there were five people involved in the process of unloading. Finally, Figure 13-c shows the unusual even of a person cleaning the dock-floor.

6.6.2 User Study For Detected Anomalies

To analyze how intuitive the detected anomalies are to humans, a user test involving 7 users was performed. First 8 regular activities for a subject were selected so she could understand the notion of a regular activity in the environment. 10 more activities were selected, 5 of which were labeled as regular by the system while the rest of the 5 were detected as anomalies. Each of the 7 users were shown these 10 activities and asked to label every one of them as a regular instance or an anomaly based on the regular activities previously shown. Each of the 10 activities were given labels based on what the majority agreed upon. 8 out of 10 activities labeled by the users, corresponded with the labels of the system. The probability of the system choosing the correct label

Fig. 13. Anomalous Activities - (a) shows a delivery vehicle leaving the loading dock with its back door still open. (b) shows an unusual number of people unloading a delivery vehicle. (c) shows a person cleaning the loading dock floor.

8 out of 10 times by chance is 4.4% [2]. This highlights the interesting fact that the anomalies detected by the proposed system fairly match the natural intuition of human observers.

6.7 Noise Analysis of n-grams in Loading Dock Environment

The results presented thus far were generated using activities with hand-labeled events. However, using low-level vision sensors to detect these events will generate noise. This invites the question as to how well would the proposed system perform over noisy data. In the following, the noise analysis to check the stability and robustness of the proposed framework is presented; allowing one to make some predictions about its performance on data using low-level vision.

Given the discovered activity-classes and the learned detection threshold using the training set of 150 activity-instances, various types and amounts of noise to the 45 test sequences was added, and the following two tests were performed:

(1) ***Regular Classification Rate:*** Percent activities classified as regular members in the 45 ground truth test activities maintain their correct activity-class and regular-membership labels in the face of noise.

(2) ***Anomaly Detection Rate:*** Percent of 45 ground truth test activities detected as anomalies still get detected as anomalies in the face of noise.

Different amounts of noise using four types of noise models, Insertion Noise, Deletion Noise, Substitution Noise and Swap Noise was synthetically generated. We generated one noisy event-symbol using a particular noise model, anywhere within a window of a time-period for each activity in the testing data set. For instance Insertion Noise of time period 10 would insert one event-symbol between any two consecutive event-symbols, every 10 symbols.

[2] Given that the probability of correctly choosing the true label by simply guessing is 0.5, the binomial probability states that chance of an 8/10 success is $C_8^{10}(0.5)^8(0.5)^2 \approx .0439$

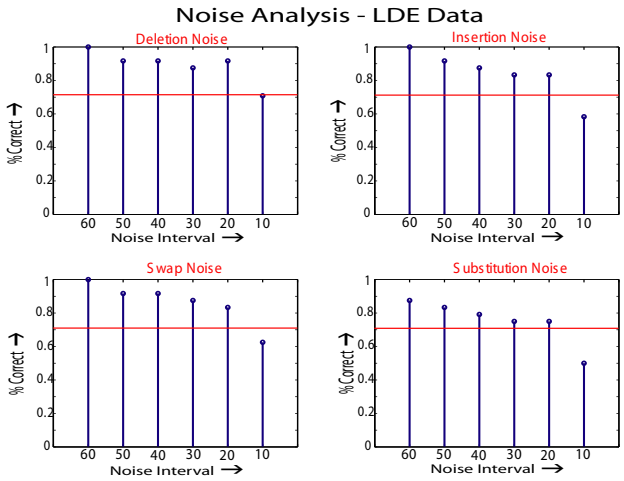


Fig. 14. Performance Analysis - Each graph shows system-performance under synthetically generated noise using different generative noise models.

The classification performance of the proposed system under such noise model is shown in Figure 14. The system performs robustly in the face of noise and degrades gracefully as the amount of noise increases. Likewise, the anomaly detection capability of our system in the face of synthetically generated noise is shown in Table **1**. The reason for such high detection rate even with large amount of synthetic noise is that it is unlikely that an anomaly would transform into something regular when perturbed randomly.

6.8 Automatic Event Detection

To move one step closer towards using low-level vision, we wrote a feature-labeling software that a user uses only to label the various objects of interest in the scene such as the doors of the loading dock, the delivery vehicles and its doors, people, packages and carts. We assign each object a unique ID during labeling. The ID numbers and object locations are stored in an XML format on a per-frame basis. We also wrote event detectors that parsed the XML data files to compute the distances between these objects for the 45 test activities. Based on the locations and velocities of these objects, the detectors performed automatic event detection.

The horizontal line in Figure 14 shows the *Regular Classification Rate* of our system over these automatically generated event sequences, *i.e.* 70.8%. The results for *Anomaly Detection Rate* for the automatically generated event sequences is 90.48%.

Noise Model	%age Correct
Insertion Noise	100%
Deletion Noise	99%
Swap Noise	97%
Substitution Noise	100%

Table 1
Anomaly Detection Rate: The average detection rate of the system in the face of noise.

6.9 Anomalous Activity Explanation

Figure 15 shows the explanation generated by the system for the three anomalous activities (shown in Figure 13). The anomaly shown in Figure 13- (a) was classified to a activity-class where people frequently carry packages through the front door of the building. There was only one person in this anomaly who delivers the package through the side door. This is evident by looking at the extraneous features of the anomaly (Figure 15-b) where the tri-gram `Person Full Handed` → `Person Exits Side Door` → `Person Empty Handed` captures this difference. The second tri-gram of Figure 15-b, `Person Full Handed` → `Person Exits Back Door` → `Person Full Handed` shows the fact that there was another person who went out of the garage to tell the driver of the delivery vehicle that his back door was open.

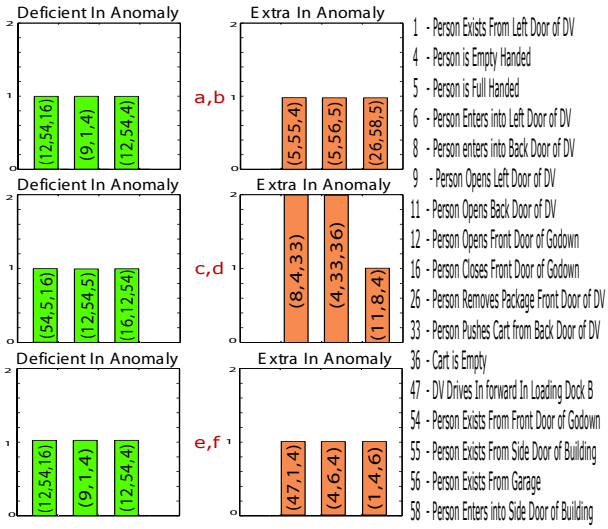


Fig. 15. Anomaly Explanation - explanations generated by the system for anomalies in Figure 13.

The membership activity-class of anomaly in Figure 13-b has people frequently carrying packages through the front door of the building. In this anomaly, all of the workers go to the side door of the building. Moreover, majority of

events in this anomaly were related to carts that is not one of the general characteristic of its membership activity-class. This is shown in Figure 15-d by tri-grams `Person Enters Back Door of DV` → `Person Empty Handed` → `Person Pushes Cart from Back Door of DV`, and `Person Empty Handed` → `Person Pushes Cart from Back Door of DV` → `Cart Empty`. Similarly Figure 15- (e) and Figure 15-f explain how anomaly in Figure 13-c was different from its membership activity-class.

7 Activity-Class Characterization

So far, we have considered situations where the beginning and end of activities is explicitly known. However, there are many scenarios where such demarcations are not so well-defined. For such situations, it is crucial to find concise characterizations of the discovered activity-classes that could be used for online activity classification and detection of anomalous activities. We formalize this problem as finding predictably recurrent activity subsequences (called event motifs) using variable-memory Markov chains (*VMMC*). Note that our usage of (*VMMC's*) in § 3.1 to empirically analyze the competence of n-grams, was for purely generative purposes. However, here we describe a novel method to learn the *VMMC* model of an activity-class in a data-driven manner.

7.1 Defining Event Motifs

We are interested in frequently occurring event subsequences that are useful in predicting future events in activities. Following [51], we assume that a class of activity sequences can be modeled as a variable-memory Markov chain (*VMMC*). We define an ***event-motif*** for an activity-class as one of the variable-memory elements of its *VMMC*. We cast the problem of finding the optimal length of the memory element of a *VMMC* as a function optimization problem and propose our objective function in the following.

7.2 Formulation of Objective Function

Let Y be the set of events, A be the set of activity-instances, and C be the set of discovered activity-classes. Let function $\mathcal{U}(a)$ map an activity $a \in A$ to its membership class $c \in C$. Let the set of activities belonging to a particular class $c \in C$ be defined as $A_c = \{a \in A : \mathcal{U}(a) = c\}$. For $a = (y_1, y_2, ..., y_n) \in A$ where $y_1, y_2, ...y_n \in Y$, let $p(c|a)$ denote the probability that activity a belongs

to class c. Then,

$$p(c|a) = \frac{p(a|c)p(c)}{p(a)} \propto \prod_{i=1}^{n} p(y_i|y_{i-1}, y_{i-2}, ..., y_1, c) \tag{20}$$

where we have assumed that all activities and classes are equally likely. We approximate Eq 20 by a *VMMC*, M_c as:

$$\prod_{i=1}^{n} p(y_i|y_{i-1}, ..., y_1, c) = \prod_{i=1}^{n} p(y_i|y_{i-1}, ..., y_{i-m_i}, c) \tag{21}$$

where $m_i \leq i-1 \ \forall \ i$. For any $1 \leq i \leq n$, the sequence $(y_{i-1}, y_{i-2}, ..., y_{i-m_i})$ is called the *context* of y_i in M_c([51]), denoted by $\mathcal{S}_{M_c}(y_i)$. We want to find the sub-sequences which can effeciently characterize a particular class, while having minimal representation in other classes. We therefore define our objective function as:

$$\mathcal{Q}(M_c|A_c) = \gamma - \lambda \tag{22}$$

where

$$\gamma \ = \ \prod_{a \in A_c} p(c|a) \ \text{ and } \ \lambda \ = \ \sum_{c' \in C \setminus \{c\}} \prod_{a \in A_{c'}} p(c'|a) \tag{23}$$

Intuitively, γ represents how well a set of event-motifs can characterize a class in terms of correctly classifying the activities belonging to that class. On the other hand, λ denotes to what extent a set of motifs of a class represent activities belonging to other classes. It is clear that maximizing γ while minimizing λ would result in the optimization of $\mathcal{Q}(M_c|A_c)$. Note that our motif finding algorithm leverages our activity-class discovery framework by using the availability of the discovered activity-classes to find the maximally mutually exclusive motifs.

7.3 Objective Function Optimization

We now explain how we optimize our proposed objective function. [51] describe a technique to compare different *VMMC* models that balances the predictive power of a model with its complexity. Let s be a context in M_c, where $s = y_{n-1}, y_{n-2}, ..., y_1$, and $y_{n-1}, y_{n-2}, ..., y_1 \in Y$. Let us define the suffix of s as $\textit{suffix}(s) = y_{n-1}, y_{n-1}, ...y_2$. For each $y \in Y$, let $N_{A'}(y, s)$ be the number of occurrences of event y in activity-sequences contained in $A' \subseteq A$ where s precedes y, and let $N_{A'}(s)$ be the number of occurances of s in activity-

sequences in A'. We define the function $\Delta_{A'}(s)$ as

$$\Delta_{A'}(s) = \sum_{y \in Y} N(s,y) log \left(\frac{\hat{p}(y|s)}{\hat{p}(y|_{suffix}(s))} \right) \tag{24}$$

where $\hat{p}(y|s) = N_{A'}(s,y)/N_{A'}(s)$ is the maximum likelihood estimator of $p(y|s)$. Intuitively, $\Delta_{A'}(s)$ represents the number of bits that would be saved if the events following s in A', were encoded using s as a context, versus having *suffix*(s) as a context. In other words, it represents how much better the model could predict the events following s by including the last event in s as part of context of these events.

We now define the function $\Psi_c(s)$ (bit gain of s) as

$$\Psi_c(s) = \Delta_{A_c}(s) - \sum_{c' \in C \setminus \{c\}} \Delta_{A_{c'}}(s) \tag{25}$$

Note that higher values of $\Delta_{A_c}(s)$ imply greater probability that an activity in A_c is assigned to c, given that s is used as a motif. In particular, higher the value of $\Delta_{A_c}(s)$, higher will be the value of γ. Similarly, higher the value of $\sum_{c' \in C \setminus \{c\}} \Delta_{A_{c'}}(s)$, higher the value of λ.

We include a sequence s as a context in the model M_c iff

$$\Psi_c(s) > K \times log(\ell) \tag{26}$$

where ℓ is the total length of all the activities in A, while K is a user defined parameter. The term $K \times log(\ell)$ represents added complexity of the model M_c, by using s as opposed to *suffix*(s) as a context, which is shorter in length and occurs at least as often as s. The higher the value of K the more parsimonious the model will be.

Equation 26 selects sequences that both appear regularly and have good classification and predictive power - and hence can be thought of as event-motifs. Work in [43] shows how the motifs in a *VMMC* can be represented as a tree. Work done in [3] presents a linear time algorithm that constructs such a tree by first constructing a data structure called a Suffix Tree to represent all subsequences in the training data A, and then by pruning this tree to leave only the sequences representing motifs in the *VMMC* for some activity-class. We follow this approach by using Equation 26 as our pruning criterion.

8 Results: Discovered Event Motifs

We now present the results of motifs we obtained using our method for the previously discovered activity-classes in Loading Dock and House environments.

8.1 Discovered Motifs for Loading Dock Environment

The highest big-gain event-motifs found for the 7 discovered activity-classes in the Loading Dock domain are:

- **Class 1:** Person places package into back door of delivery vehicle → Person enters into side door of building → Person is empty handed → Person exists from side door of building → Person is full handed → Person places package into back door of delivery vehicle.
- **Class 2:** Cart is full → Person opens front door of building → Person pushes cart into front door of building → Cart is full → Person closes front door of building → Person opens front door of building → Person exists from front door of building → Person is empty handed → Person closes front door of building.
- **Class 3:** DV drives in forward into LDA → Person opens left door of DV → Person exists from left door of DV → Person is empty handed → Person closes the left door of delivery vehicle.
- **Class 4:** Person opens back door of DV → Person removes package from back door of DV → Person removes package from back door of DV → Person removes package from back door of DV → Person removes package from back door of DV → Person removes package from back door of DV.
- **Class 5:** Person closes front door of building → Person removes package from cart → Person places package into back door of DV → Person removes package from cart → Person places package into back door of DV → Person removes package from cart → Person places package into back door of DV.
- **Class 6:** Person Removes Cart From Back Door of DV → Person Removes Package From Back Door of DV → Person Places Package Into Cart → Person Places Package Into Cart → Person Removes Package From Back Door of DV → Person Places Package Into Cart → Person Removes Package From Back Door of DV → Person Places Package Into Cart.
- **Class 7:** Person closes back door of DV → Person opens left door of DV → Person enters into left door of DV → Person is empty handed → Person closes left door of DV.

8.2 Discovered Motifs - House Environment

The highest big-gain event-motifs found for the 5 discovered activity-classes in the House scenario are given below:

- **Class 1:** Alarm → Kitchen entrance → Fridge → Sink → Garage door (inside).

- **Class 2:** Stairway → Fridge → Sink → Cupboard → Sink.
- **Class 3:** Stairway → Dining Table → Den → Living-room Door → Sun-room → Living-room door → Den.
- **Class 4:** Den → Living-room door → Den → Kitchen Entrance → Stairway.
- **Class 5:** Fridge → Dining Table → Kitchen Entrance → Fridge → Sink

The discovered motifs of activity-classes seem to characterize these classes efficiently. Note that the discovered motifs for activity-classes where package delivery occurred, have events like Person Places Package In The Back Door Of Delivery Vehicle and Person Pushes Cart In The Front Door of Building→ Cart is Full. On the other hand event-motifs for activity-classes where package pick-up occurred, have events such as Person Removes Package From Back-Door Of Delivery Vehicle and Person Places Package Into Cart. Similarly, The motifs for the House environment capture the position where the person spends most of her time and the order in which she visits the different places.

8.3 Subjective Assessment of Discovered Motifs

Given some data, our proposed discovery method would, by construction, find some motifs for the discovered activity-classes. This begs the question as to how could one ascertain the veracity of our results. Since our final goal is to design a system that would be able to discover and characterize human-interpretable activity-classes, we performed a limited user test involving 7 participants, to subjectively assess the performance of our system. For each participant, 2 of the 7 discovered activity classes were selected from the Loading Dock environment. Each participant was shown 6 example activities, 3 from each of the 2 selected activity-classes. The participants were then shown 6 motifs, 3 for each of the 2 classes, and were asked to associate each motif to the class that it best belonged to. Their answers agreed with our systems 83% of the time, *i.e.*, on average a participant agreed with our system on 5 out of 6 motifs. The probability of agreement on 5 out of 6 motifs by random guessing [3] is only 0.093.

9 Conclusions & Future Work

In this work, we presented a novel activity representation of event n-grams for unsupervised activity analysis in sensor-rich environments. We have specifically investigated if there is sufficient structural signature at a local tempo-

[3] According to the binomial probability function the chance of randomly agreeing on 5 out of 6 motifs is $C_5^6(0.5)^1(0.5)^5$.

ral scale that can entail a reasonably disjunctive partitioning of the activity space. Using this representation we investigated unsupervised activity discovery, activity-class characterization, and anomalous activity explanation in a maximally informative manner. Results over extensive data-sets, collected from multiple sensor-rich environments are presented, to show the competence of our framework.

Our proposed representation of activities as bags of event n-grams is an attempt towards extracting activity structure from local event-statistics, as opposed to having to script the different ways in which activities can take place. Such an approach can be helpful in large uncontrolled settings where explicitly encoding the activity structure is infeasible. While the proposed activity-representation captures both content and order information of events in activities, it does pose the problem of dealing with sparse high dimensional data. It is evident that higher values of n would capture the temporal order information of events more rigidly, and would entail a more discriminative representation. This discriminative power would however come at the cost of an exponential growth in the dimensionality of the space. The exact value of n used for a particular environment is a function of the event dynamics for that environment.

Making use of this representation, we have shown how different activity-classes can be discovered in an unsupervised manner, by exploiting the notion of maximal cliques in edge-weighted activity-graphs. The empirical analysis of the discovered activity-classes in both the examined active-settings, shows that the cohesiveness of the discovered classes is dependent on how semantically meaningful the event vocabulary is. The granularity at which one chooses to define this vocabulary offers a tradeoff between the prior knowledge needed to describe the events, and their expressiveness. Dictated by the activity-dynamics and the physical attributes of the environment, event vocabulary reflects the initial bias that we induce in the system. The nature and amount of this bias must be chosen judiciously.

Activity class discovery is generally a precursor to anomaly detection. Our current framework models anomalies simply as being different from the regular activities. While this approach allows us to detect potentially anomalous activities without having to define them *a priori*, it does result in relatively higher false-positive rate. One way of reducing this high false-positive rate is to show the detected anomalies to a human observer, based on whose feedback learn a better estimate of what is meant by being anomalous in an environment. We leave this question as an open problem for future work.

The idea of explaining why an activity is anomalous, can be helpful in large scale surveillance systems. While we believe that the selection of the key features to explain the detected anomalies is a step in the right direction, much

remains to be done in terms of learning the dependence of these features amongst one another and how that can effect the interpretability of such explanations. Moreover, in certain cases, the presence or absence of just one event in an activity, can go a long way in explaining why that activity is anomalous. Incorporating such factors in the current framework of anomalous activity explanation remains an open problem for the future.

As there are numerous situations where the start and end demarcations of activity instances are not so well-defined, it is crucial to find concise characterizations of the discovered activity-classes that could be used for online activity classification and detection of anomalous activities. Our formalization of this problem as finding predictably recurrent event motifs using variable-memory Markov chains (*VMMC*) is an efficient means to this end. Our initial user study seems to suggest that the discovered motifs succinctly capture the discriminative structures of the discovered activity-classes.

References

[1] D. Aha, D. Kibler, and M. Albert. Instance-based learning algorithms. *Journal of Machine Learning*, 6, 1991.

[2] J. F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26:832–843, 1983.

[3] A. Apostolico and J. Bejerano. Optimal amnesic probabilistic automata. *Journal of Computational Biology*, 7:381–393, 2000.

[4] J. Auguston and J. Miker. An analysis of some graph theoretical clustering techniques. *Journal of ACM*, 17(4):571–588, 1970.

[5] T. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proc International Conference of Intelligent Systsystems in Molecular Biology, pp. 28-36*, 1994.

[6] G. Bejerano and G. Yona. Modeling protein families using probabilistic suffix trees. In *In the Proc. of International Conference of Research in Computational Molecular Biology*, 1999.

[7] A. Bobick. Movement, activity and action: the role of knowledge in the perception of motion. In *Movement, Activity and Action: the Role of Knowledge in the Perception of Motion, Royal Society Workshop on Knowledge-based Vision in Man and Machine.*, 1997.

[8] A.F. Bobick, S.S. Intille, J.W. Davis, F. Baird, C.S. Pinhanez, L.W. Campbell, Y.A. Ivanov, A. Schuette, and A. Wilson. The kidsroom: A perceptually-based interactive and immersive story environment. In *Vismod*, 1996.

[9] M. Brand, N. Oliver, and A. Pentland. Coupled hidden markov models for complex action recognition. In *IEEE Conference of Computer Vision and Pattern Recognition*, 1997.

[10] M. Chen, T. Kanade, D. Pomerleau, and Henry Rowley. Anomaly detection through registration. In *Pattern Recognition 32(1): 113-128*, 1999.

[11] M. T. H. Chi. Conceptual change within and across ontological categories. *Cognitive models of Science: Minnesota Studies in Philosophy of Science*, pages 129–186, 1992.

[12] B. Chiu, E. Keogh, and S. Lonardi. Probabilistic discovery of time series motifs. In *SIGKDD*, 2003.

[13] T. Choudhury, M. Philipose, D. Wyatt, and J. Lester. Towards activity databases: Using sensors and statistical models to summarize people's lives. In *IEEE Data Engineering Bulletin*, 2006.

[14] T. Choudhury, J. Rehg, V. Pavlovic, and A. Pentland. Boosting and structure learning in dynamic bayesian networks for audio-visual speaker detection. In *In Proceedings of IEEE ICPR*, 2002.

[15] A. Dey, R. Hamid, C. Beckmann, I. Li, and D. Hsu. a cappella: programming by demonstration of context-aware applications. In *SIGCHI*, pages 33–40, 2004.

[16] R. Diestel. *Graph Theory (Graduate Texts in Mathematics)*. Springer; 2 edition, 2000.

[17] P. Gardenfors and K. Holmqvist. Concept formation in dimensional spaces. *Lund University Cognitive Studies*, (26), 1994.

[18] W. E. Grimson. The combinatorics of local constraints in model-based recognition and localization from sparse data. *Journal of the ACM*, 33(4):658–686, 1986.

[19] D. Gusfield. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press; 1st edition, 1997.

[20] R. Hamid, A. Johnson, S. Batta, A. Bobick, C. Isbell, and G. Coleman. Detection and explanation of anomalous activities: Representing activities as bags of event n-grams. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.

[21] R. Hamid, S. Maddi, A. Johnson, A. Bobick, I. Essa, and C. Isbell. Discovery and characterization of activities from event-streams. In *International Conference of UAI*, 2005.

[22] M. Hammouda and M. S. Kamel. Efficient phrase-based document indexing for web document clustering. *IEEE Trans. on KDE*, 16(10):1279–1296, 2004.

[23] I. Heller and C. Tompkins. An extension of a theorem of dantzig's. In *Linear Inequalities and Related Systems*, page 247254. Princeton University Press, 1956.

[24] S. Hongeng, F. Bremond, and R. Nevatia. Representation and optimal recognition of human activities. In *In Proceedings of IEEE CVPR*, 2000.

[25] S. Hongeng and R. Nevatia. Multi-agent event recognition. In *In Proc. of IEEE CVPR*, 2001.

[26] K. Ilgun, R. Kemmerer, and P. Porras. State transition analysis: A rule-based intrusion detection approach. *IEEE Transaction on software engineering*, pages 188–199, 1995.

[27] Y. Ivanov and A. Bobick. Recognition of visual activities and interactions by stochastic parsing. *Pattern Analysis and Machine Intelligence Journal*, 22(8):852–872, 2000.

[28] A. Johnson and A. Bobick. Relationship between identification metrics: Expected confusion and area under a roc curve. In *In Proceedings of IEEE CVPR*, 2002.

[29] J. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46, 1999.

[30] J. Kleinberg. An impossibility theorem for clustering. In *In Proceedings of 16th conference on Neural Information Processing Systems*, 2002.

[31] C. Largeron. Prediction suffix trees for supervised classification of sequences. *Pattern Recognition Letters*, 2003.

[32] W. Lee and S. Stolfo. A framework for constructing features and models for intrusion detection systems. *ACM Transactions on Information and System Security*, 3(4), November 2000.

[33] L. Liao, D.J. Patterson, D. Fox, and H. Kautz. Learning and inferring transportation routines. *Artificial Intelligence. J.*, 2007.

[34] R. Mann, A. Jepson, and J. Siskind. The computational perception of scene dynamics. *Proceedings of Fourth ECCV*, 1996.

[35] L. Manor and M. Irani. Event-based video analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2001.

[36] D. Minnen, I. Essa, and T. Starner. Expectation grammars: Leveraging high-level expectations for activity recognition. In *IEEE Conference on CVPR. Madison, WI.*, 2003.

[37] T. Oates. Peruse: An unsupervised algorithm for finding recurring patterns in time series. In *IEEE ICDM, Japan.*, 2002.

[38] N. Oliver, E. Horvitz, and A. Garg. Layered representations for human activity recognition. In *IEEE ICMI*, 2002.

[39] M. Pavan and M. Pelillo. A new graph-theoretic approach to clustering and segmentation. In *IEEE Conference of CVPR*, 2003.

[40] R. Polana and R. Nelson. Low level recognition of human motion. *IEEE Workshop on Non-rigid and Articulated Motion*, 1994.

[41] L. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *In Proceedings of the IEEE*, pages 257–286, 1989.

[42] V. Raghavan and C. Yu. A comparison of the stability characteristics of some graph theoretic clustering methods. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 3:393–402, 1981.

[43] D. Ron, Y. Singer, and N. Tishby. The power of amnesia: learning probabilistic automata with variable memory length. *Machine Learning*, 25:117–149, 1996.

[44] E. Rosch, C.Mervis, W. Gray, D. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8, 1976.

[45] G. Salton. *The SMART Retrieval System - Experiment in Automatic Document Processing*. Prentice-Hall, Englewood Cliffs, New Jersey, 1971.

[46] R. Schank. *Dynamic Memory: A Theory of Reminding and Learning in Computers and People*. Cambridge University Press, 1983.

[47] C. Stauffer and W. Grimson. Learning patterns of activity using real-time tracking. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 22(8):747–757, 2000.

[48] G. Sukthankar and K. Sycara. Robust recognition of physical team behaviors using spatio-temporal models. In *Proceedings of the fifth international joint conference on Autonomous agents and multiagent systems*, pages 638–645, 2006.

[49] P. Tse, J. Intriligator, J. Rivest, and P. Cavanagh. Attention and the subjective expansion of time. *Perception and Psychophysics*, 66:1171–1189, 2004.

[50] S. Ullman. *The Interpretation of Visual Motion*. MIT Press, 1979.

[51] M. Weinberger, J. Rissanen, and M. Feder. A universal finite memory source. In *IEEE Trans. Inform. Theory, vol. IT-41, pp. 643–652, 48*, 1995.

[52] A. Yuille and N. Grzywacz. A computational theory for the perception of coherent visual motion. *Nature*, 333:71–74, may 1988.

[53] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. In *In Proc. of IEEE CVPR*, 2004.