Augmenting Bag-of-Words: Data-Driven Discovery of Temporal and Structural Information for Activity Recognition

Vinay Bettadapura[1] vinay@gatech.edu
Grant Schindler[1] schindler@gatech.edu
Thomas Plötz[2] thomas.ploetz@ncl.ac.uk
Irfan Essa[1] irfan@cc.gatech.edu

[1]Georgia Institute of Technology, Atlanta, GA, USA [2]Newcastle University, Newcastle upon Tyne, UK
http://www.cc.gatech.edu/cpl/projects/abow

Abstract

We present data-driven techniques to augment Bag of Words (BoW) models, which allow for more robust modeling and recognition of complex long-term activities, especially when the structure and topology of the activities are not known a priori. Our approach specifically addresses the limitations of standard BoW approaches, which fail to represent the underlying temporal and causal information that is inherent in activity streams. In addition, we also propose the use of randomly sampled regular expressions to discover and encode patterns in activities. We demonstrate the effectiveness of our approach in experimental evaluations where we successfully recognize activities and detect anomalies in four complex datasets.

1. Introduction

Activity recognition in large, complex datasets has become an increasingly important problem. Extracting activity information from time-varying data has applications in domains such as video understanding, activity monitoring for healthcare and surveillance. Traditionally, sequential models like Hidden Markov Models (HMMs) and Dynamic Bayesian Networks have been used to address activity recognition as a time-series analysis problem. However, the assumption of Markovian dynamics restricts the application of such sequential models to relatively simple problems with known spatial and temporal structure of the data to be analyzed [22]. Similarly, syntactic methods like Parse Trees and Stochastic Context Free Grammars [18, 11] are not well suited for recognizing weakly structured activities and are not robust to erroneous or uncertain data.

As a promising alternative, research in activity recognition from videos and other time-series data has moved towards bag-of-words (BoW) approaches and away from the traditional sequential and syntactic models. However, while BoW approaches are good at building powerful and sparser representations of the data, they completely ignore the ordering and structural information of the particular words regarding their absolute and relative positions. Furthermore, standard BoW approaches do not account for the fact that different types of activities have different temporal signatures. Each event in a long-term activity has a temporal duration, and the time that passes between each pair of consecutive events, is different for different activities

We introduce novel BoW techniques and extensions that explicitly encode the temporal and structural information gathered from the data. Recent activity recognition approaches such as [19] have extended the BoW approach with topic models [23] using probabilistic Latent Semantic Analysis [10] and Latent Dirichlet Allocation [1], leading to more complex classification methods built on top of standard BoW representations. In contrast, we increase the richness of the features in the BoW representation and with the use of standard classification backends (like k-NN, HMM and SVM), we demonstrate that our augmented BoW techniques lead to better recognition of complex activities.

Contributions: We describe a method to represent temporal information by quantizing time and defining new temporal events in a data-driven manner. We propose three encoding schemes that use n-grams to augment BoW with the discovered temporal events in a way that preserves the local structural information (relative word positions) in the activity. This narrows the conceptual gap between BoW and sequential models. In addition, to discover the global patterns in the data, we augment our BoW models with randomly sampled Regular Expressions. This sampling strategy is motivated by the random subspace method as it is commonly used for decision tree construction [2] and related approaches which have shown success in a wide variety of classification and visual recognition problems [14].

We evaluate our approach in comparison to standard BoW representations on four diverse classification tasks: *i)* Vehicle activity recognition from surveillance videos (Section 4.1); *ii)* Surgical skill assessment from surgery videos

(Section 4.2); *iii)* Unsupervised learning of player roles in soccer videos (Section 4.3) and *iv)* Recognition of human behavior and anomaly detection in massive wide-area airborne surveillance (simulation) data (Section 4.4). Recognition using our augmented BoW outperforms the standard BoW approaches in all four datasets. We provide evidence that this superior performance generalizes to any classification framework by demonstrating how sequential models (HMMs), instance based learning (k-NNs), and discriminative recognition techniques (SVMs) benefit from the new representation and outperform respective models trained on standard BoW. Finally, we show how augmented BoW-based techniques successfully unveil further details of the analyzed datasets, such as behavior anomalies.

2. Related Work

The Bag of Words (BoW) model was first introduced for Information Retrieval (IR) with text [21]. Since then, it has been used extensively for text analysis, indexing and retrieval [16]. Building on the success of BoW approaches for IR with text and images, research in activity recognition has focused on working with BoW built using local spatio-temporal features [25] and more recently with robust descriptors, which exploit continuous object motion and integrate it with distinctive appearance features [4], features based on dense trajectories [24] and features learnt in an unsupervised manner directly from video data [13].

While the focus has mostly been on recognizing human activities in controlled settings, recent BoW based approaches have focused on recognizing human activities in more realistic and diverse settings [12], and with the use of higher level semantic concepts (attributes) that allow for more descriptive models of human activities [15]. However, when activities are represented as bags of words, the underlying sequential information provided by the ordering of the words is typically lost. To address this problem, n-grams have been used to retain some of the ordering by forming sub-sequences of n items [16] (Figure 2). More recently, variants of the n-gram approach have been used to represent activities in terms of their local event sub-sequences [9]. While this preserves local sequential information and causal ordering, adding absolute and relative temporal information results in more powerful representations as we demonstrate in this paper.

Our augmentation method is independent of the underlying BoW representation, i.e., the modality of the data to be processed. The input to our algorithm is a sequence of atomic events, i.e., words. On video data these can be either derived from state-of-the-art short-duration event detectors (e.g., the Actom Sequence Model [7], automatic action annotation [5]), or any other suitable feature detectors.

3. Activity Recognition with Augmented BoW

We define an *activity* as a finite sequence of events over a finite period of time where each *event* in the activity is an occurrence. For example, if "start", "turn", "straight" and "stop" are four individual events, then a vehicle driving activity will be a finite sequence of those events over some finite time (e.g. "start $\rightarrow$ straight$\rightarrow$ turn$\rightarrow$ stop$\rightarrow$ start$\rightarrow$ straight$\rightarrow$ stop"). We call these events, that can be described by an observer and have a semantic interpretation, as *observable events*.

Recent methods for activity recognition try to detect such observable events and build BoW upon it. However, the temporal structure underlying the activities that shall be recognized is typically neglected. The time taken by each observable event and the time elapsed between two subsequent events are two important properties that contribute to the temporal signature of an activity that is being performed. For example, a car at a traffic light will have a shorter time gap between the "stop" and "start" events than a delivery vehicle that has to stop for a much longer time (until its contents are loaded/unloaded) before it can start again.

3.1. Discovering Temporal Information

We represent activities as sequences of discrete, observable events. Let $\omega = \{a_1, a_2, a_3, \ldots, a_p\}$ denote a set of p activities, and let $\phi = \{e_1, e_2, e_3, \ldots, e_q\}$ denote the set of q types of observable events. Each activity a_i is a sequence of elements from ϕ. Each event type can occur multiple times at different positions in a_i.

We now introduce *temporal events*. Let $\tau_{j,k}$ be the temporal event defined as the time elapsed between the end of observable event e_j and the start of observable event e_k, where $k > j$. Since it measures time, $\tau_{j,k}$ is non-negative. Also, let $\pi_{j,k}$ be the temporal event defined as the time elapsed between the start of observable event e_j and the end of observable event e_k, where $k \geq j$. Thus, $\tau_{j,k}$ measures the time elapsed between any two events whereas $\pi_{j,k}$ measure the time elapsed between any two events including the time taken by those two events. Thus, $\tau_{j,k}$ and $\pi_{j,k}$ are related by the equation $\pi_{j,k} = \pi_{j,j} + \tau_{j,k} + \pi_{k,k}$. We posit that these two types of temporal events, $\tau_{j,k}$ and $\pi_{j,k}$, can model all the temporal properties of an activity. The four possible scenarios are listed here:

1. $\tau_{j,j+1}$: Time elapsed between any two consecutive events e_j and e_{j+1}
2. $\tau_{j,k}$: Time elapsed between any two events e_j and e_k, where $k > j$
3. $\pi_{j,j}$: Time taken by a single event e_j
4. $\pi_{j,k}$: Time taken by set of events e_j to e_k, where $k \geq j$

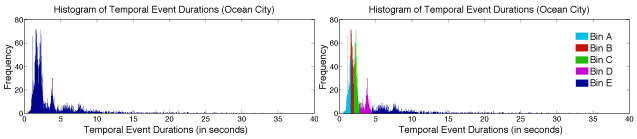

Figure 1. Histogram of event durations for Ocean City dataset (left) and data-driven creation of temporal bins (right; $N = 5$).

To work with these temporal events, we will have to quantize them into a finite number of N bins. This quantization is crucial in allowing us to incorporate a notion of time into BoW models. However, uniformly dividing the time-line into N bins is not ideal. As illustrated by the temporal event duration histograms of $\tau_{j,j+1}$ for the Ocean-City dataset (see Section 4.1) in Figure 1, short and medium duration temporal events occur much more frequently than longer duration temporal events. Similar temporal distributions are observed in the other datasets we have analyzed.

To ensure that we capture the most useful temporal information, we pursue a data-driven approach for binning. Bins are selected based on the distribution of temporal events. If there are S temporal events, then we divide the temporal space into N bins such that each of the N bins contains an equal proportion S/N of the temporal events (illustrated in Figure 1 for $N = 5$). Note that, if the time-line had been naively divided into 5 equally sized bins, then most of the temporal events would have been placed in the first bin while the other 4 bins would have been almost empty. The choice of N depends on the problem we are addressing. Lower values of N result in increased loss of temporal information.

Example 1: Say, temporal event $\tau_{j,k}$ is of 4 second duration and temporal event $\tau_{l,m}$ is of 20 second duration, then from Figure 1, we see that $\tau_{j,k}$ will be assigned to bin D and $\tau_{l,m}$ will be assigned to bin E. Let ψ denote the function that maps the temporal events to their respective temporal bins. Then, we can say that $\psi(\tau_{j,k}) = D$ and $\psi(\tau_{l,m}) = E$.

There are many possible ways by which we can encode these new temporal events along with the observable events to build augmented BoW representations. The simplest way would be to just add the quantized temporal events to the BoW, i.e., if the BoW contained x observable events and we extracted y new quantized temporal events, then the augmented BoW will now contain $x + y$ number of elements. Although this naive representation already gives better results than just the BoW (see Section 4), as shown in the next section, more sophisticated alternatives are possible.

3.2. Encoding Local Structure

In the following we describe three encoding schemes we have developed that merge the temporal events with the observable events in a way that captures local structure.

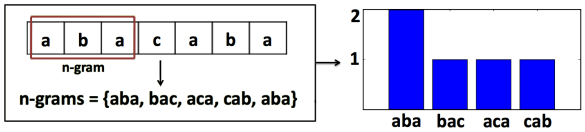

Figure 2. Building n-grams and their histogram (here $n = 3$) [9]

3.2.1 Interspersed Encoding

In interspersed encoding, the main focus is on the time elapsed between every pair of consecutive events. Let $\tau_{j,j+1}$ be a temporal event defined as the time elapsed between any two consecutive observable events e_j and e_{j+1} in activity a_i. Once the quantized temporal events $\psi(\tau_{j,j+1})$ are computed for all event pairs $e_j, e_{j+1} \in a_i$, they are then inserted into a_i at their appropriate positions between events e_j and e_{j+1}. Let this new sequence of *interspersed* events for activity a_i be denoted by T_i. In general, if activity a_i has d events, then after the inclusion of the quantized temporal events, T_i will have $2d-1$ events (the original d observable events plus the new $d-1$ temporal events).

Example 2: For the activity $a_1 = (e_1, e_2, e_3)$, we have $T_1 = (e_1, \psi(\tau_{1,2}), e_2, \psi(\tau_{2,3}), e_3)$. If temporal event $\tau_{1,2}$ is of 4 second duration and $\tau_{2,3}$ is of 20 second duration, then the quantized temporal events will be $\psi(\tau_{1,2}) = D$ and $\psi(\tau_{2,3}) = E$. So, the interspersed sequence of events for activity a_1, will be $T_{1=}(e_1, D, e_2, E, e_3)$.

One of the main drawbacks of classical BoW representations is the loss of original word orderings (i.e. local structural information). This is particularly adverse in the context of activity recognition because activities correspond to causal chains of observable and temporal events. Losing the ordering will result in a loss of all causality and contextual information. We employ n-grams in order to retain ordering of events [6]. An n-gram is a sub-sequence of n terms from a given sequence. Deriving n-grams and their histograms from a given sequence is illustrated in Figure 2.

Using this approach, for every activity a_i, the event sequence T_i is transformed into an n-gram sequence T_i^I (where the superscript I stands for *interspersed*). This T_i^I feature vector representing activity a_i is the final result of interspersed encoding. From Example 2, with $n = 3$, the event sequence $T_{1=}(e_1, D, e_2, E, e_3)$ will be transformed into the n-gram sequence $T_1^I = (e_1De_2, De_2E, e_2Ee_3)$ or in its histogram form $T_1^I = \{e_1De_2 \Rightarrow 1, De_2E \Rightarrow 1, e_2Ee_3 \Rightarrow 1\}$ (denoted as key-value pairs where the key is the n-gram and the value is its frequency).

3.2.2 Cumulative Encoding

In cumulative encoding, the main focus is on the cumulative time taken by a subsequence of observable events. Let $\psi(\pi_{j,j+n-1})$ be a quantized temporal event defined as the total time taken by n consecutive events e_j to e_{j+n-1} in ac-

tivity a_i. Once the quantized temporal event $\psi(\pi_{j,j+n-1})$ is computed for the consecutive sequence of observable events $e_j \dots e_{j+n-1} \in a_i$, it is appended to the set of the observable events. Let this new sequence of "cumulative" observable and temporal events for activity a_i be denoted by T_i^C (where the superscript C stands for "cumulative").

Example 3: If activity $a_2 = (e_1, \dots, e_5)$, n = 3, then $T_2^C = (e_1e_2e_3\psi(\pi_{1,3}), e_2e_3e_4\psi(\pi_{2,4}), e_3e_4e_5\psi(\pi_{3,5}))$. Say, $\pi_{1,3}$ is of 4 second duration, $\pi_{2,4}$ is of 20 second duration and $\pi_{3,5}$ is of 1 second duration and that $\psi(\pi_{1,3}) = D$, $\psi(\pi_{2,4}) = E$ and $\psi(\pi_{3,5}) = A$. So, the new sequence of events for activity a_2, will be $T_2^C = (e_1e_2e_3D, e_2e_3e_4E, e_3e_4e_5A)$ or in histogram form $T_2^C = \{e_1e_2e_3D \Rightarrow 1, e_2e_3e_4E \Rightarrow 1, e_3e_4e_5A \Rightarrow 1\}$.

Interspersed encoding focuses on the time elapsed between events whereas cumulative encoding focuses on the time taken by the events.

3.2.3 Pyramid Encoding

Given the choice of encoding scheme —either interspersed or cumulative— in pyramid encoding all l-grams of length $l, \forall l \in [1, n]$ are generated. Then we build a pyramid of these l-grams allowing for processing of event sequences at multiple scales of resolution. We denote BoW representations for activity a_i generated through pyramid encoding by T_i^P.

The output of each of these encoding schemes, i.e., T_i^I, T_i^C and T_i^P is the augmented BoW model containing the observable and temporal events, encoded in a way that captures the local structure.

3.3. Capturing Global Structure

While n-grams are good at capturing local information, their capability to capture longer range relationships are rather limited. This is where regular expressions come into play. Obviously, it is computationally intractable to enumerate all possible regular expressions for a given vocabulary of observable and temporal events. Thus, given the set of observable events ϕ and the set of discovered temporal events N, we construct a vocabulary of all events $\phi \cup N$ denoted by Γ where $|\Gamma| = |\phi| + |N|$, and create a sub-space of regular expressions by restricting their form to:

$$\wedge \ .* \ (\alpha) \ (\beta_1 \mid \dots \mid \beta_r)\varphi \ (\gamma) \ .* \ \$ \qquad (1)$$

where the symbols $\alpha, \beta_i, \gamma \in \Gamma$ with $i \in [1, r]$ and $r = \text{rand}\,(1, |\Gamma|)$. The symbol φ is randomly set to one of the three quantifier characters: $\{*, +, ?\}$. The special characters have the following meaning: " $\wedge$ " matches the start of the sequence, "." matches any element in the sequence, " $*$ " matches the preceding element zero or more times, " $+$ " matches the preceding element one or more times, "?" matches the preceding element zero or one time and "$" matches the end of the sequence. The "|" operator matches either of its arguments. For example, $e_1(e_2|e_3)e_4$ will match either $e_1e_2e_4$ or $e_1e_3e_4$.

The first symbol that will be matched (α) and last symbol that will be matched (γ) are chosen randomly from Γ using probability-proportional-to-size sampling (PPS) and the r intermediate symbols β_i are chosen randomly from Γ using simple random sampling (SRS). PPS concentrates on frequently occurring events and picks the first and last symbols in the regular expression to be the ones that have the greatest impact on the population estimates whereas SRS chooses each of the intermediate symbols with equal probability, thus giving a fair chance for all events to equally participate in the matching process. The results of our experimental evaluation suggest that this combination of PPS-SRS sampling of the regular expression subspace strikes the right balance between discovering global patterns across activities and discovering the anomalous activities.

Regular expressions of the above form are randomly generated and those that do not match at least one of the activities/event-sequences are rejected. Accepted regular expressions are treated as new words and added to our augmented BoW representation. This final representation now contains automatically discovered temporal information and both local and global structural information of the activities. Our experiments show that increasing the number of words in BoW through randomly generated regular expressions by just 20% boosts the activity recognition and anomaly detection results significantly (Section 4).

3.4. Activity Recognition

Activity recognition using augmented BoW is pursued in a straightforward manner by feeding the time-series data in their novel representation into statistical modeling backends. Note that there is in principle no limitation on the kind of classification framework to be employed. In Section 4 we present results for instance based learning (k-NN), sequential modeling (HMM), and discriminative modeling (SVM).

Given videos or time-series data of activities, temporal information is discovered using the histogram method described in Section 3.1. Using n-grams, the temporal information is then merged with the extracted BoW thereby preserving local ordering of the words. The new BoW model is then further augmented by adding new words created using randomly sampled regular expressions (to capture global patterns in the data), and then processed by the statistical modeling backend for actual activity recognition.

4. Experimental Evaluation

The methods presented in this paper were developed in order to improve BoW-based activity recognition, thereby aiming for generalization across application domains. For practical validation, we have thus evaluated our approaches

Figure 3. Sample frame from Ocean City data showing the various objects being tracked.

in a range of experiments that cover three diverse classes of learning problems (binary classification, multi-class classification, and unsupervised learning) across four challenging datasets from different domains.

Optimization of the estimation procedure for augmented BoW representations involves the two main parameters in our system: N, the number of temporal bins used for quantization and n, the size of the n-gram used for encoding. Low values of N and n result in the loss of temporal and structural information whereas high values can lead to large BoW with very high dimensionality. The optimal values for N and n are determined by standard grid search [3]. Within a user supplied interval, all grid points of (N,n) are tested to find the combination that gives the highest accuracy. 50% of the particular datasets is held-out for parameter optimization, and the remaining 50% is used for model estimation using cross-validation. This provides an unbiased estimate of the generalization error and prevents over-fitting.

The main evaluation criterion for all activity recognition experiments is classification accuracy, which we report as absolute percentages and, for more detailed analysis, in confusion matrices. For the first set of experiments (Section 4.1) we compare three different classification backends (k-NNs with cosine-similarity distance metric, HMMs, and SVMs) and explore their capabilities in systematic evaluations of their parameter spaces. Due to space constraints the presentation of results for the remaining set of experiments is limited to those achieved with the k-NN classification backend. These results are, however, representative for all three types of classifiers evaluated.

k-NNs with cosine-similarity distance metric, i.e. Vector Space Models (VSM), treat the derived BoW vectors of activities as document vectors and allow for automatic analysis in terms of querying, classifying, and clustering the activities [16]. Prior to classification, each term in our augmented BoW is assigned a weight based on its term-frequency and document-frequency in order to obtain a statistical measure of its importance. Classification is done using leave-one-out cross-validation (LOOCV).

HMM-based experiments employ semi-continuous modeling with Gaussian mixture models (GMM) as feature space representations [6]. GMMs are derived by means of an unsupervised density learning procedure. All HMMs are

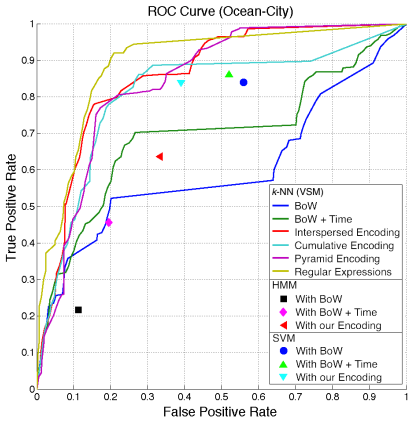

Figure 4. Classification results for Ocean City dataset. Our encoding schemes outperform the BoW baseline on three classification backends: VSM, sequential models (HMMs) and SVMs.

based on linear left-right topologies with automatically derived model lengths (based on training data statistics), and are trained using classical Baum-Welch training. Classification is pursued using Viterbi-decoding. Parameter estimation and model evaluation employs 10-fold cross-validation.

Experiments on SVMs are carried out in 10-fold cross-validation using LIBSVM with an RBF kernel. Parameter optimization utilizes a grid-search procedure as it is standard for finding optimal values for C and γ [3].

4.1. Ocean City Surveillance Data

The first dataset consists of 7 days of uncontrolled videos recorded at Ocean City, USA [20]. The input video was stabilized and geo-registered and $2,140$ vehicle tracks were extracted using background subtraction and multi-object tracking [20] (Figure 3). An event detector analyzed the tracks, detected changes in structure over time and represented each track by a sequence of observable events. The types of events detected in each track were "start", "stop", "turn" and "u-turn".

Out of the $2,140$ vehicle tracks, 448 vehicles are either entering or exiting parking areas on either side of the road (Figure 3). The recognition objective is to determine whether or not vehicles are involved in parking activities.

With the empirically determined optimal values of $N = 2$ and $n = 2$, we perform binary classification. The results are shown in Figure 4. For k-NN based experiments, ROC curves were generated by varying the acceptance threshold. Augmenting BoW with temporal information (bag-of-words + time) improves the results over the BoW baseline. The performance is further improved with our proposed Interspersed, Cumulative and Pyramid encoding schemes. However, the best results are obtained when we augment our BoW with randomly generated regular expressions.

Figure 4 also shows the performance of HMM and SVM based recognition backends using augmented BoW


Figure 5. Long-range (left) and close-up (right) stills of video footage from training sessions for surgical skill assessment. Participants practice suturing using regular instruments and suture pads.

representations. Both techniques produce fixed decisions based on maximizing models' posterior probabilities, i.e., no threshold-based post-processing is applied for the actual recognition. Consequently, ROC curves are not applicable, and the particular results are shown as points in the figure.

Analyzing the evaluation results, it becomes evident that: *i)* our proposed encoding schemes outperform the BoW baseline; and *ii)* superior classification accuracy generalizes across recognition approaches (k-NN, HMM, SVM), with largest gain for Vector Space Models.

4.2. Surgical Skill Assessment

The second set of experiments is related to evaluating surgical skills as it is standard routine in practical training of medical students. As part of a larger case-study, 16 medical students were recruited to perform typical suturing activities (stitching, knot tying, etc.) using regular instruments and tissue suture pads. Both long-range and close-up videos of these "suturing" procedures were captured at 50 fps at a resolution of 720p (sample still images in Figure 5). As part of the training procedure, participants completed 2 sessions with 2 attempts in each session, resulting in a total of 64 videos. Ground truth annotation was done by an expert surgeon who assessed the skills of the participants using a standardized assessment scheme (OSATS [17]) based on 7 different metrics (Table 1) on a three-point scale (low competence, medium, and high skill).

Harris3D detectors and histogram of optical-flow (HOF) descriptors [25] are used to extract visual-words from the surgery videos. BoW are built with vocabularies constructed using k-means clustering (with $k = 50$), and then augmented using our techniques. Table 1 summarizes our experiments (using k-NN classification backend) and gives comparisons with the BoW baseline. It can be seen that augmented BoW based approaches outperform the BoW baseline in all 7 skill metrics with an overall accuracy of 72.56%.

Since our augmented BoW representations capture time and co-occurrence of words, we hypothesized that an automated analysis procedure using augmented BoW should perform particularly well in assessing the "time and motion" and "knowledge of procedure" skills. Recognition results reported in Table 1 indicate that this is indeed the case. The classification accuracies are 74.60% and 80.95% (an increase of 23.81% and 20.63% respectively, over the BoW baseline), thus validating our hypothesis.

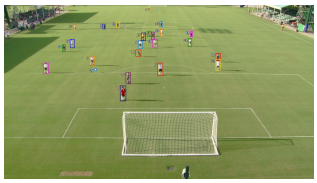
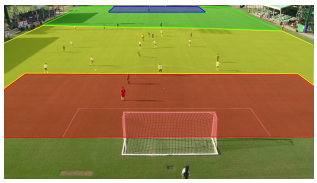

Figure 6. Sample stills from soccer videos dataset. **Left:** The 24 objects being tracked: 22 players from both teams, referee and the ball. **Right:** The 4 zones used by our event detector: Zone-A (Red), Zone-B (Yellow), Zone-C (Green) and Zone-D (Blue).

	RI	ARI	NMI
BOW baseline	0.7984	0.2922	0.6147
BOW + Time	0.8300	0.3920	0.6974
Our Encoding	**0.8261**	**0.5244**	**0.7462**

Table 2. Cluster quality on soccer videos dataset. The 3 metrics used are Rand Index (RI), Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). Our encoding (Interspersed encoding with 3-grams, 3 time bins and 20 random regular expressions) gives better cluster quality than the BoW baseline.

4.3. Learning Player Activities from Soccer Videos

Automatic detection, tracking and labeling of the players in soccer videos is critical for analyzing team tactics and player activities. Previous work in this area has mostly focussed on detecting and tracking the players, recognizing the team of the players using appearance models and detecting short-duration player actions. In our experiments, we consider the problem of unsupervised learning of long-range activities and roles the various players take on the field. Given their tracks, we cluster them into 7 clusters: "Team-A-Goalkeeper", "Team-A-Striker", "Team-A-Defense", "Team-B-Goalkeeper", "Team-B-Striker", "Team-B-Defense" and "Referee".

We analyzed full length match videos (720p at 59.94 fps) from the Disney Research soccer games dataset and tracked the 24 objects (players, referee, and ball) on the field using a multi-agent particle filter based framework [8] (Figure 6). The tracks were given to an event detector that divided the field into 4 zones (Figure 6) and detected 10 types of events: "Enter-Zone-A", "Leave-Zone-A", "Enter-Zone-B", "Leave-Zone-B", "Enter-Zone-C", "Leave-Zone-C", "Enter-Zone-D", "Leave-Zone-D", "Receive-Ball" and "Send-Ball". With this vocabulary of 10 events, we built augmented BoW and clustered them using k-means clustering where $k = 7$. Clustering results are given in Table 2. It can be seen that augmented BoW outperform the BoW baseline on all 3 cluster quality metrics. In a supervised setting, we achieve an accuracy of 82.61%, which is a 17.39% improvement over the BoW baseline (which is 65.22%).

	Respect for tissue	Time and motion	Instrument handling	Suture handling	Flow of operation	Knowledge of procedure	Overall performance	Average accuracy
M1: BOW baseline	66.67%	50.79%	50.79%	69.84%	49.21%	60.32%	52.38%	57.14%
M2: BOW + Time	69.84%	66.67%	65.08%	69.84%	63.49%	74.60%	68.25%	68.25%
M3: Our encoding	**73.02%**	**74.60%**	**68.25%**	**73.02%**	**66.67%**	**80.95%**	**71.43%**	**72.56%**

Table 1. Surgical skill assessment using OSATS assessment scheme [17]. Ground truth annotation provided by an expert surgeon who assessed the training sessions using 7 different metrics (columns) and a three-point scale (low competence, medium, and high skill). Results given are accuracies from automatic recognition using k-NN, replicating expert assessment based on video footage of the training sessions. Our encoding (Interspersed encoding with 3-grams, 5 time bins and with 20 random regular expressions) outperforms the BoW baseline on all 7 metrics.

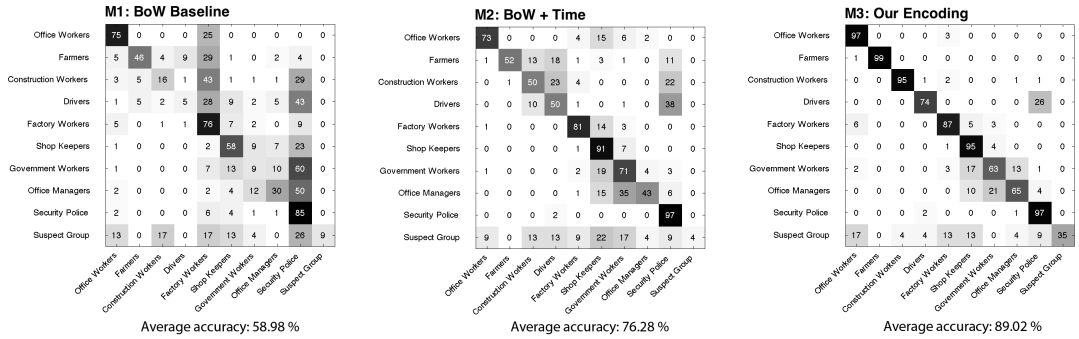

Figure 7. Results on WAAS dataset: **Left:** BoW baseline; **Middle:** BoW + Time; **Right:** Our encoding (5-grams, 5 time bins and with $1,000$ random regular expressions). Overall improvement of 30.04% is observed with our method (compared to standard BoW baseline).

4.4. Wide Area Airborne Surveillance (WAAS)

In order to evaluate the applicability and scalability of our approach on massive datasets with several hundreds of thousands of activities, we consider the Wide-Area Airborne Surveillance (WAAS) simulation dataset.

The WAAS dataset was developed by the U.S. Military as part of their Activity Based Intelligence (ABI) initiative. The goal is to capture motion imagery from an airborne platform that provides persistent coverage of a wide area, such as a town or a small city, and merge the automatically captured data from the aerial station with intelligence gathered by ground forces to build a surveillance database of humans and vehicles in that area. In order to aid research in this area, the WAAS dataset has been released, which contains Monte Carlo simulation of the activities of $4,623$ individuals for a total duration of 46.5 hours generated in 1 minute increments. There are a total of 180 events (like "Eat Lunch", "Enter Vehicle", "Exit Vehicle", "Move", "Wait", etc) with a total of $544,777$ event sequences spread across $28,682$ buildings. Ground truth labels are available on the 10 different professions of all the individuals. 23 out of the $4,623$ individuals are suspected to be part of a terror group.

Given this large database, we show that our augmented BoW can successfully classify people's professions and detect some of the suspect individuals based on the temporal and structural similarities in their activities. Classification accuracies and confusion matrices are shown in Figure 7. Note that, with our encoding, more than a third of the suspect group are correctly classified which baseline methods failed to capture. This successful identification of suspicious behavior is especially remarkable since those suspects aim for imitating "normal" behavior and thus their activities are very similar to harmless activities.

4.5. Test for Statistical Significance

With McNemar's chi-square test (with Yates' continuity correction), we check for the statistical significance between the results of our two multi-class classification problems (Figure 7 and Table 1). For the surgery dataset, though all the 7 skill classifications were statistically significant, due to space constraints, only results on "knowledge of procedure" classification is presented.

The null hypothesis is that the improvements are due to chance. However, as shown in Table 3 for both the datasets, the χ^2 values are greater than the critical value (at 95% significance level) of 3.84 and the p-values are less than the significance level (α) of 0.05. Thus, the null hypothesis can be rejected and we can conclude that the improvements obtained with our methods are statistically significant.

5. Conclusion

BoW models are a promising approach to real-world activity recognition problems where only little is known a-priori about the underlying structure of the data to

	M1 vs M2	M1 vs M3
χ^2	165.09	530.35
p-value	< 0.0001	0.0026

	M1 vs M2	M1 vs M3
χ^2	4.76	9.33
p-value	0.0291	0.0023

Table 3. McNemar's tests on statistical significance between the different methods on the 2 multi-class classification problems. Each column compares two methods. **Left:** Comparing the methods in Figure 7 for the WAAS dataset; **Right:** Comparing the methods in Table 1 for the "knowledge of procedure" skill in the surgery dataset (the other 6 skill classifications were also statistically significant, but are not shown due to space constraints).

be analyzed. We presented a significant extension to BoW-based activity recognition, where we augment BoW with temporal information and with both local and global structural information, using temporal encoding, n-grams and randomly sampled regular expressions, respectively. In addition to generally improved activity recognition, our approach also detects anomalies in the data, which is important, for example in human behavior analysis applications. We have demonstrated the capabilities of our approach on both real-world vision problems and on massive wide-area surveillance simulations.

Acknowledgements Funding and sponsorship was provided by the U.S. Army Research Office (ARO) and Defense Advanced Research Projects Agency (DARPA) under Contract No. W911NF-11-C-0088 and W31P4Q-10-C-0262. Parts of this work have also been funded by the RCUK Research Hub on Social Inclusion through the Digital Economy (SiDE), and the German Research Foundation (DFG, Grant No. PL554/2-1). Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of our sponsors.

We thank Kitware, Disney Research and DARPA for providing the Ocean City, Soccer and WAAS datasets, respectively.

References

[1] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *JMLR*, 2003. 1

[2] Leo Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996. 1

[3] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Trans. on Intelligent Systems and Technology*, 2011. 5

[4] M. Chen, L. Mummert, P. Pillai, A. Hauptmann, and R. Sukthankar. Exploiting multi-level parallelism for low-latency activity recognition in streaming video. In *ACM SIGMM Conf. on Multimedia Systems*, 2010. 2

[5] O. Duchenne, I. Laptev J., Sivic, F. Bach, and J. Ponce. Automatic annotation of human actions in video. In *ICCV*, pages 1491–1498, 2009. 2

[6] G. A. Fink. *Markov Models for Pattern Recognition, From Theory to Applications*. Springer, 2008. 3, 5

[7] A. Gaidon, Z. Harchaoui, and C. Schmid. Actom sequence models for efficient action detection. In *CVPR*, 2011. 2

[8] R. Hamid, R. K. Kumar, M. Grundmann, K. Kim, I. Essa, and J. Hodgins. Player localization using multiple static cameras for sports visualization. In *CVPR*, 2010. 6

[9] R. Hamid, S. Maddi, A. Johnson, A. Bobick, I. Essa, and C. Isbell. A novel sequence representation for unsupervised analysis of human activities. *Artificial Intell.*, 173:1221–44, 2009. 2, 3

[10] T. Hofmann. Probabilistic latent semantic indexing. In *ACM SIGIR Conf. on IR*, 1999. 1

[11] Y. A. Ivanov and A. F. Bobick. Recognition of visual activities and interactions by stochastic parsing. *PAMI*, 22(2):852–872, August 2000. 1

[12] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 2

[13] Q. Le, W. Zou, S. Yeung, and A. Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *CVPR*, 2011. 2

[14] V. Lepetit and P. Fua. Keypoint recognition using randomized trees. *PAMI*, 28(9):1465–1479, 2006. 1

[15] J. Liu, B. Kuipers, and S. Savarese. Recognizing human actions by attributes. In *CVPR*, 2011. 2

[16] C. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge Univ. Press, 2008. 2, 5

[17] JA Martin, G. Regehr, R. Reznick, H. MacRae, J. Murnaghan, C. Hutchison, and M. Brown. Objective structured assessment of technical skill (osats) for surgical residents. *British Journal of Surgery*, 84(2):273–278, 1997. 6, 7

[18] Darnell Moore and Irfan Essa. Recognizing multitasked activities from video using stochastic context-free grammar. In *AAAI*, 2002. 1

[19] J. Niebles, H. Wang, and Li Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. *IJCV*, 2008. 1

[20] S. Oh, A. Hoogs, M. Turek, and R. Collins. Content-based retrieval of functional objects in video using scene context. In *ECCV*, 2010. 5

[21] G. Salton. *The SMART retrieval system: Experiments in automatic document processing*. Prentice-Hall, 1971. 2

[22] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. Machine recognition of human activities: A survey. *IEEE Trans. Circuits Systems for Video Tech.*, October 2008. 1

[23] H.M. Wallach. Topic modeling: beyond bag-of-words. In *ICML*, pages 977–984. ACM, 2006. 1

[24] H. Wang, A. Klaserr, C. Schmid, and C. Liu. Action recognition by dense trajectories. In *CVPR*, 2011. 2

[25] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009. 2, 6