Exploiting Typicality for Selecting Informative and Anomalous Samples in Videos

Jawadul H. Bappy, *Student Member, IEEE*, Sujoy Paul, *Student Member, IEEE*, Ertem Tuncel, *Senior Member, IEEE,* and Amit K. Roy-Chowdhury, *Fellow, IEEE*

***Abstract*—In this paper, we present a novel approach to find informative and anomalous samples in videos exploiting the concept of typicality from information theory. In most video analysis tasks, selection of the most informative samples from a huge pool of training data in order to learn a good recognition model is an important problem. Furthermore, it is also useful to reduce the annotation cost as it is time-consuming to annotate unlabeled samples. Typicality is a simple and powerful technique which can be applied to compress the training data to learn a good classification model. In a continuous video clip, an activity shares a strong correlation with its previous activities. We assume that the activity samples that appear in a video form a Markov chain. We explicitly show how typicality can be utilized in this scenario. We compute an atypical score for a sample using typicality and the Markovian property, which can be applied to two challenging vision problems- (a) sample selection for learning activity recognition models, and (b) anomaly detection. In the first case, our approach leads to a significant reduction of manual labeling cost while achieving similar or better recognition performance compared to a model trained with the entire training set. For the latter case, the atypical score has been exploited in identifying anomalous activities in videos where our results demonstrate the effectiveness of the proposed framework over other recent strategies.**

***Index Terms*—Activity Recognition, Typicality, Sample Selection, Active Learning, Anomaly and Novelty Detection.**

I. Introduction

Classification tasks such as activity recognition, rely on labeled data in order to learn a recognition model. With an increase in the availability of visual data, it is a manually intensive job to continuously label them. In [1], it has been shown that more labeled data does not always help a recognition model to learn better; sometimes the performance might even degrade due to noisy data points. Thus, selection of the most informative samples to train a recognition model becomes crucial. Furthermore, automatic detection of unusual or abnormal activities is an area of significant interest in diverse video analysis applications. We address both these problems in this paper. We present *an information-theoretic approach for obtaining a subset of informative samples to learn a good classification model for activity recognition, and for identifying anomalous/irregular activities in videos.*

In computer vision, the selection of informative samples [2] has been widely used to reduce the manual labeling effort for annotation tasks and to train a good recognition model. Most of the sample selection methods devise a sample-wise informativeness utility score based on which the samples are selected for manual labeling [2], [3], [4]. However, they are highly dependent on classifier uncertainty or diversity in the feature space. Furthermore, the aforementioned approaches consider the individual samples to be independent. Recent works [5], [6], [7], [8] exploit the inter-relationships (or contextual information) between samples in order to reduce the number of labeled samples to train the recognition models with applications including activity recognition, scene and object classification, document classification, etc. Most of these approaches involve graphical models to exploit the interrelationships between the samples, where inference and joint entropy computation becomes intractable in the case of acyclic graphs and requires simplifying assumptions. Moreover, these methods introduce high computational complexity at the inference step as the number of nodes increases.

The analysis of abnormal activities in videos has been of growing interest in security and surveillance applications. Most of the anomaly detection methods [9], [10], [11] train a model to learn the patterns of normal activities and consider an activity as abnormal whose pattern is deviated from the normal activities. Some methods [12], [13], [14] exploit local statistics of low-level features, local spatio-temporal descriptors, and bag-of-word approach to detect anomalies in videos. Recent efforts [11], [15] in anomaly detection consider interrelationship between the activities in identifying abnormal activities. In [15], temporal regularity patterns are learned from the normal activities in order to detect unusual activity. In this paper, we introduce a new way of measuring the irregularity by utilizing temporal relation-

ship between activities to detect abnormal activities in video. The abnormal/anomalous activities are excluded from the training phase. Thus, the task of identifying anomalous sample can be referred as novelty detection.

In information theory, the idea of 'typical set' has been one of the core tools behind several data processing applications such as data compression, data transmission, data security, and data search. It is based on the intuitive notion that not all the messages are equally important, i.e., some messages carry more information than others. By analogy, we can exploit this concept to reduce the manual labeling cost by choosing the most informative samples from a large pool of unlabeled data. Typicality allows representation of any sequence using entropy as a measure of information [16]. The concept of typical sets is developed on the basis of the asymptotic equipartition property (AEP) which is analogous to the law of large numbers in probability theory. Consider a sequence $x_1, \ldots, x_n$ of i.i.d. random variables with probability mass function $p(x)$. According to the AEP, as n approaches infinity, the *empirical entropy* of the sequence converges to the actual entropy $H(X)$ for distribution $p(x)$. For finite n, the set of all sequences x^n is divided into two sets, the typical set, where the empirical entropy is close to the actual entropy, and the atypical set. Collectively, the typical set has a very high probability, and all sequences in it have roughly the same probability that can be approximated as $2^{-nH(X)}$ for high n. (please see Sec. III for more details).

This notion of typicality can be utilized for informative subset selection, with the labels or a group of labels of samples being a random variable. A sequence not belonging to the typical set (atypical) may be termed as informative as it does not follow the distribution of the random variable learned from the previously labeled instances. For example, in activity recognition, different activities may be temporally connected, e.g., a person opening a car trunk followed by the person carrying an object. If a different set of semantic entities appear in a particular scene, then the atypical score, computed based on the deviations in typicality, would be high and will be identified as informative. Thus, the natural interactions between semantic entities can prove to be a rich source of information in order to identify informative samples for applications like active learning, and anomaly detection.

Our previous work [7] showed preliminary results employing the concept of typicality on joint classification tasks, e.g, scene-object for images. In this work, we extend this idea more thoroughly for a range of computer vision problems in videos, such as selection of informative samples for training a model, and anomaly detection in videos. Moreover, [7] employs typicality

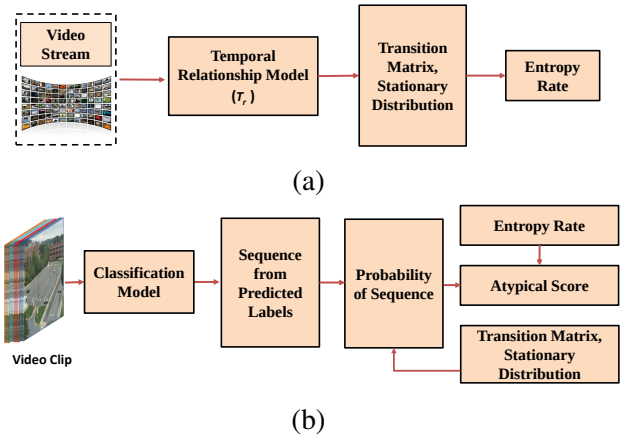


Fig. 1: The figures present how typical set can be applied in vision problems to compute an atypical score. (a) represents training phase where we learn entropy rate by computing transition matrix and stationary distribution (please see Sec. III-B). In (b), a sequence is generated from the prediction of activity labels given a test video. Then, the probability of the sequence is calculated from the transition matrix and stationary distribution which are learned during the training phase. Finally, we compute an atypical score for each sample in a video.

by utilizing information flow from scene or activity to objects in a joint classification scenario, by conditioning on the former. However, it does not deal with the inter-relationships between activities, which we study in this paper. In activity recognition, the current activity may be strongly correlated with the previous activity sample, and can be represented as Markovian. In this paper, we assume that action samples produce a Markov chain where the current sample only depends on the previous sample, and demonstrate how to utilize typicality for this scenario. We design an utility function which depends on the length of a sequence (please see Sec. III for more details). Moreover, we show that typicality based sample selection approach is computationally faster than existing graph-based approaches [6], [8], [5] that exploit the correlation between the samples. From the experimental results, we observe that proposed approach outperforms other state-of-the-art methods by large margin to reduce the manual labeling cost. The atypical score can also be applied to detect abnormal activities in videos, which will be discussed later in Sec. III-D2.

A. *Framework Overview*

Fig. 1 presents the overview of our proposed method. We can divide the overall process into two phases: (a) training phase, and (b) testing phase.

Activities in a video are represented as a Markov chain where the current activity depends on previous activity only. During the training phase, we learn the recognition model ($\mathcal{M}$) and the temporal relationship model (T_r). T_r could be a simple co-occurrence statistic that captures the correlation between two consecutive activities. We learn transition matrix and stationary probability using this temporal co-occurrence. We compute entropy rate required to define a typical set, details of which are provided in Sec. III.

At test phase, a video clip is fed into the classifier $\mathcal{M}$. $\mathcal{M}$ provides predicted labels with a confidence score. We form a sequence from the predicted labels obtained from $\mathcal{M}$ and compute uncertainty (please see details in Sec. III) of the sequence. We compare this uncertainty with the entropy of source distribution obtained from T_r in order to compute the atypical score. We can also calculate entropy from the distribution of predicted scores for each sample using $\mathcal{M}$. With this uncertainty score and atypical score, we formulate an optimization function to choose the most informative set of samples to be labeled manually by a human annotator. We also used the atypical score to detect anomalies in videos.

We applied the proposed approach to two applications- (a) informative sample selection, and (b) anomaly detection. For the first scenario, we present our approach from the perspective of batch mode active learning, where the goal is to select the most informative samples to update the recognition model in an online setting where unlabeled data are coming continuously in batches. By solving the optimization function mentioned above, we can find informative samples which will be considered for manual labeling. With these newly labeled samples, $\mathcal{M}$ and T_r are updated. In this process, we intend to achieve similar performance with the model which is trained on all the samples (100% manual labeling). In anomaly detection, we consider whole training set to understand the nature of normal activities. We learn the typical model and recognition model. Given a test sample, we set a threshold on atypical score to determine whether an activity sample is abnormal or not.

Contributions: Our *major contributions* are as follows.

- In this paper, we show how the concept of typicality in information theory can be applied to different computer vision problems, namely (a) activity recognition, and (b) anomaly detection in videos.
- Unlike [7], where the variables in a sequence are independent, we show how the concept of 'typical set' can be applied to temporally dependent variables in computer vision problems. We demonstrate our strategy on videos instead of images as presented in [7].
- We perform rigorous experimentation on two scenarios- (1) sample selection for activity classification, (2) detection of abnormal activities. Our framework on sample selection outperforms state-of-the-art methods significantly in reducing the manual labeling cost while achieving same recognition performance compared with a model trained on all the samples. We also demonstrate the usefulness of the method in finding anomalies in videos.

II. Related Work

In this section, we will briefly discuss the related work on visual recognition task, sample selection, anomaly detection, and typicality.

Visual Recognition Task. The proposed framework applies to work in activity classification. Some promising approaches in computer vision use context model [17], [5] on top of recognition model in order to achieve higher accuracy. In [5], spatio-temporal relationship and co-occurrence statistics have been utilized in order to recognize activities in video. Most of the context based approaches exploit conditional random field (CRF) to interrelate the samples, which become computationally expensive as nodes in the graph increases. Recently, various deep learning based models have been presented in [18], [19], [20], [21], [22] for activity classification. These frameworks show promising performance in recognizing activities.

Sample Selection Methods. Some of the state-of-the-art sample selection approaches are expected change in gradients [2], information gain [3], expected prediction loss [4], and expected model change [23] to obtain the samples for querying. Some of the common techniques to measure uncertainty for selecting the informative samples are presented in [24], [4]. Along with classifier uncertainty, diversification in the chosen samples is introduced by using k-means [25] or sparse representative subset selection [26]. In [25], the authors incorporated two strategies - best vs. second best and K-centroid to select the informative subset. The aforementioned approaches consider the individual samples to be independent. Recent advances [6], [8], [5] in active learning incorporate contextual relationships to reduce manual labeling cost without compromising recognition performance. Most of these approaches involve graph-based models where the belief is propagated through nodes using inference algorithm. These approaches might be computationally expensive as the number of nodes increases.

Anomaly Detection. Several works [11], [27] have exploited semantically meaningful activities in order to detect anomalies. A comprehensive review of anomaly

detection is provided in [9]. [10] presents a hierarchical framework for identifying local and global anomalies utilizing hierarchical feature representation and Gaussian process. In [28], the authors present a method that exploits Locality Sensitive Hashing Filters (LSHF), which hashes normal activities into multiple feature buckets. [29] proposes a space-time Markov Random Field (MRF) model to identify abnormal activities in videos. Some works [30], [11] exploit spatio-temporal context in order to detect anomalous activities. In [31], the authors present an approach that learns both dominant and anomalous behaviors in videos of different spatio-temporal complexity. In anomaly detection, deep learning based approaches such as sparse auto-encoder [32] and fully convolutional feed-forward network [15], [33] are also utilized. The paper [34] exploits unmasking in order to detect abnormal events in a video.

Typicality. The concept of 'typical set' [35] has widely been used in various applications like data compression, data transmission, data security, and data search [36], [37], [38]. In [39], the authors define atypicality as the deviation of the information from average. Then, it is applied in universal source coding and a number of real world datasets. In computer vision, the term 'typicality' is mentioned in some research papers for several tasks such as category search [40], object recognition [41], and scene classification [42]. However, they do not exploit the notion of information-theoretic *typical set*. In [7], a novel active learning method was proposed exploiting the theory of *typical set*. In this paper, we extend the work presented in [7] by demonstrating its generalizability across a variety of computer vision problems with a special focus on the dynamics of human activities.

III. Typicality and Its Application in Videos

In information theory, a typical set represents a set of sequences drawn from an i.i.d distribution, whose total probability of occurrence is close to one. A sequence can be categorized into either typical or atypical, depending on whether it belongs to the typical set or not. There are two kinds of typicality, namely, weak and strong. In this problem, we focus on weak typicality to develop our sample selection framework. Next, we will briefly show the concept of weak typicality and then, demonstrate how typicality can be used in different computer vision tasks.

A. Typicality in Information Theory

Let us consider $\boldsymbol{x}^n$ to denote a sequence $x_1, \ldots, x_n$ drawn from an i.i.d distribution $P_{X^n}(.)$, whose empirical entropy can be expressed as,

$$\begin{aligned} -\frac{1}{n}\log_2 P_{X^n}(\boldsymbol{x}^n) &= -\frac{1}{n}\log_2 \prod_{i=1}^{n} P_{X_i}(x_i) \\ &= -\frac{1}{n}\sum_{i=1}^{n}\log_2 P_{X_i}(x_i) \quad (1) \end{aligned}$$

By the weak law of large numbers Eqn. 1 can be written as

$$-\frac{1}{n}\sum_{i=1}^{n}\log_2 P_{X_i}(x_i) \rightarrow E[-\log_2 P_{X^n}(\boldsymbol{x}^n)] = H(X) \quad (2)$$

Definition. A set of sequences with probability distribution $P_{X^n}(.)$ can be considered as weakly typical set if it satisfies the following criteria:

$$\Big| -\frac{1}{n}\log_2 P_{X^n}(\boldsymbol{x}^n) - H(X) \Big| \leq \epsilon \quad (3)$$

Let us denote $\mathcal{E} = -\frac{1}{n}\log_2 P_{X^n}(\boldsymbol{x}^n) - H(X)$, which represents atypical score of a sequence. Next, we will demonstrate how this typical set [16] concept can be exploited to compute atypical score for Markov chain.

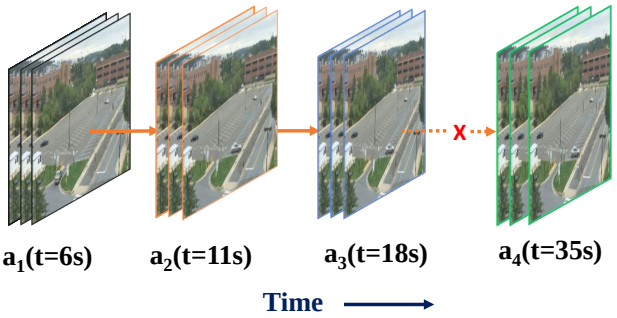


Fig. 2: The figure presents how different activities appear over time in a video. The appearance of current activity depends on the previous activity as marked by right arrow. However, this dependency (temporal context) is determined by the time interval between two activities. For instance, the temporal link between the activities - a_3 and a_4, is discarded due to the long time interval. For all other cases, the temporal connection between activity samples is established as illustrated in the figure.

B. Asymptotic Equipartition Property for Markov Chain

In this section, we will show how to compute the atypical score for a Markov chain, motivated by the assumption that sequential activities exhibit Markovian property. We aim to exploit the Asymptotic Equipartition Property (AEP) for Markov Chain in computer vision problem. This has been a well-established theorem

[43], [44] applied to several other domains such as data compression, and data transmission. Fig. 2 shows an example of different activities in a video that are connected via a temporal link. We can assume this temporal ordering in terms of Markov chain, where current activity only depends on previous activity. Let us consider a stochastic process, where states can be denoted as $\{X_1, X_2, \ldots, X_n\}$ and each state $X_i \in \mathcal{X}$. If a source $X_1, X_2, \ldots X_n$, produces a sequence, we can characterize the distribution of a sequence as $P\{(X_1, \ldots, X_n) = (x_1, \ldots, x_n)\} = P(x_1, \ldots x_n)$. Since we assume the temporal link as Markov chain, we can write the conditional independence as follows.

$$P(x_{n+1}|x_n, \ldots, x_1) = P(x_{n+1}|x_n) \tag{4}$$

In Markov chain, one state moves successively to next state with a probability. Let us denote current state X_i which moves to next state X_j with probability p_{ij}. The probability p_{ij} is called transition probability. In activity recognition, we assume that the transition probability does not change over time. So, the Markov chain becomes time-invariant (stationary) where the conditional probability $P(x_{n+1}|x_n)$ does not rely on n. This can be written as

$$P_{X_1 \ldots X_n}(x_1, \ldots, x_n) = P_{X_{1+t} \ldots X_{n+t}}(x_1, \ldots, x_n). \tag{5}$$

Here, t denotes time shift. For stationary Markov chain, we can define a transition matrix T_s, where each entry represents the probability of jump from one state to another. The transitional matrix T_s can be written as

$$T_s = \begin{bmatrix} p_{11} & p_{12} & p_{13} & . & p_{1n} \\ p_{21} & p_{22} & p_{23} & . & p_{2n} \\ . & . & . & . & . \\ p_{n1} & p_{n2} & p_{n3} & . & p_{nn} \end{bmatrix}$$

where, each $p_{ij} \geq 0$ and for all states X_i,

$$\begin{aligned} \sum_{m=1}^{n} p_{im} &= \sum_{m=1}^{n} p(x_m|x_i) \\ &= 1. \end{aligned} \tag{6}$$

Now, we can compute the probability of a sequence, $P(X_1, X_2, \ldots, X_q)$ as

$$P(X_1, X_2, \ldots, X_q) = P(X_1)P(X_2|X_1) \ldots P(X_q|X_{q-1}) \tag{7}$$

Here, q is the number of elements in a sequence. If the Markov chain is stationary, then we can define a stationary distribution μ over all X_i. The stationary distribution can be computed as

$$\mu = \mu T_s \tag{8}$$

where, each element of μ would be $\mu_i = \sum_{j=1}^{n} \mu_j p_{ji}$, and $\sum_{i=1}^{n} \mu_i = 1$. If we transpose Eqn. 8, we obtain

$$\begin{aligned} (\mu T_s)^\intercal &= \mu^\intercal \\ T_s^\intercal \mu^\intercal &= \mu^\intercal \end{aligned} \tag{9}$$

Thus, stationary distribution can be obtained from the eigenvector of $T_s^\intercal$ with eigenvalue 1 by utilizing eigen value decomposition. If the transition matrix T_s is known, we can compute stationary distribution. It could be possible to have multiple eigenvectors associated to an eigenvalue of 1 where each eigenvector gives rise to an associated stationary distribution. In this case, the Markov chain becomes reducible, i.e. has multiple communicating classes [45].

Entropy Rate: The entropy rate of a stochastic process $\{X_1, X_2, \ldots, X_n\}$ can be written as

$$H(\mathcal{X}) = \lim_{n \to \infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n) \tag{10}$$

For stationary process, the entropy rate [46] becomes

$$H(\mathcal{X}) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1). \tag{11}$$

$H(X_n|X_{n-1}, \ldots, X_1)$ is non-increasing as n increases and the limit must exist [46]. For Markov chain, the above equation 11 would be $H(\mathcal{X}) = \lim_{n \to \infty} H(X_n|X_{n-1}, \ldots, X_1) = \lim_{n \to \infty} H(X_n|X_{n-1})$. If $X_i \sim \mu$, then the entropy rate is

$$H(\mathcal{X}) = -\sum_{ij} \mu_i p_{ij} \log p_{ij} \tag{12}$$

Using Eqns. 9 and 12, we can compute stationary distribution and entropy rate. From the Asymptotic Equipartition Property (AEP) theorem, the probability of a sequence (Eqn. 7) becomes

$$p(X_1, \ldots, X_q) \to 2^{-qH(\mathcal{X})}. \tag{13}$$

Now, we denote the atypical score of a sequence as $\mathcal{E} = -\frac{1}{q} \log p(X_1, \ldots, X_q) - H(\mathcal{X})$. Next, we will demonstrate how this atypical score can be utilized in a couple of applications- (a) sample selection and (b) anomaly detection.

C. Computation of Atypical Score in Video Applications

In activity classification, the next activity is related to the last activity as well as the preceding activities. So, the activity samples form a higher order of dependency than the dependency in a Markov chain. Even though a high-order stochastic process model is suitable to represent the high-order dependency, a high-order stochastic process model is undesirable due to its model

complexity and computational cost. It becomes even harder with a large dataset. Furthermore, determining the accurate order of dependency to represent the temporal relations is impractical. Thus, we have used a Markov chain instead of a high-order stochastic process model which approximately captures the temporal behavior between the activity samples. There are some works [47], [48], [49], [50] which exploit Markov model or hidden Markov model for the activity recognition task.

An activity sample is a video snippet of multiple frames and contains the category of an activity. A video clip contains a number of activity samples where the activities appear sequentially over time. An activity might be strongly correlated with its previous activity, and it is also possible that two consecutive activities are uncorrelated. Thus, we consider that a temporal link is established if the time interval between current and previous activities is below a threshold δ_t, else it is possible that two consecutive activities are not temporally related. Fig. 2 shows an example of such scenario. From the figure, we can see that temporal link between last two activities is not established due to a long time interval. In this paper, we assume that the current activity only depends on previous activity, thus $p(a_n|a_{n-1},\ldots,a_1) = p(a_n|a_{n-1})$.

As the activities form a sequence and generate Markov chain for a video clip, we can compute atypical score by computing entropy rate and the probability distribution of a sequence. For transition matrix, we simply count the frequency of an activity occurring given the previous activity. Consider, i^{th} row vector $\boldsymbol{r}^i$ of matrix T_s shown in Eqn. 14, which can be written as

$$\boldsymbol{r}^i = \frac{1}{\sum_{k=1}^{N_a}\phi_k^i}[\phi_1^i,\ldots,\phi_n^i]. \tag{14}$$

Here, N_a represents the number of activity classes. ϕ_k^i implies the number of times activity class a_k occurs given previous activity class a_i Thus, each (i,k)-th entry of $\boldsymbol{r}^i$ represents the transitional probability from a_i to a_k. After obtaining the transition matrix, we can easily compute stationary probability μ_a using Eqn. 9 by utilizing eigen value decomposition. Let us define i^{th} state $A_i \in \mathcal{A}$, which may have an outcome among the activity labels $a_1,\ldots,a_{N_a}$. So, we can compute the entropy rate as follows.

$$H(\mathcal{A}) = -\sum_{ij}\mu_{ai}\boldsymbol{r}_j^i \log \boldsymbol{r}_j^i \tag{15}$$

Given a video, set of activity samples $A_1,\ldots,A_q$ form an activity sequence. q is the number of activities occurring in a video. The probability of an activity sequence can be presented as $P(A_1,\ldots,A_q) = P(A_1)P(A_2|A_1)\ldots P(A_q|A_{q-1}) = P(\mathcal{A}^q)$. It can be calculated from μ_a and $\boldsymbol{r}_j^i$. We can compute the atypical score of the sequence as follows.

$$\mathcal{E} = -H(\mathcal{A}) - \frac{1}{q}\log_2 P(\mathcal{A}^q) \tag{16}$$

Now, in order to compute the atypical score for each of the samples, we remove a tuple from the sequence associated with i^{th} activity sample a_i and observe the deviation of atypical score as similar to Eqn. 16. An activity sample at time t (a^t) might be excluded if it appears very far from activities before and after it. In an extreme case, we only compute the entropy of that activity sample, which will be discussed in Sec. III-D. If we remove i^{th} sample from the sequence, then we compute the probability of a new sequence as $P(\mathcal{A}^{q'}) = P(A_1)P(A_2|A_1)...P(A_{i-1}|A_{i-2})P(A_{i+2}|A_{i+1})...P(A_q)$. Thus, $P(A_i|A_{i-1})$ and $P(A_{i+1}|A_i)$ are eliminated from the distribution function. The length of new sequence would be $q-2$. So, the atypical score of new sequence would be $\mathcal{E}_i$, which can be written as

$$\mathcal{E}_i = -H(\mathcal{A}) - \frac{1}{q-2}\log_2 P(\mathcal{A}^{q-2}) \tag{17}$$

We can now measure the deviation between $\mathcal{E}$ and $\mathcal{E}_i$.

$$\begin{aligned}\tilde{\mathcal{E}}_i &= |\mathcal{E} - \mathcal{E}_i| \\ &= \left|-\frac{1}{q}\log_2 P(\mathcal{A}^q) + \frac{1}{q-2}\log_2 P(\mathcal{A}^{q-2})\right| \\ &= \left|\frac{2}{q(q-2)}\sum_{\substack{m=1 \\ m\neq\{i-1,i\}}}^{q}\log_2 P(A_{m+1}|A_m) - \right. \\ &\qquad \left.\frac{1}{q}\log_2 P(A_i|A_{i-1}) - \log_2 P(A_{i+1}|A_i)\right| \end{aligned} \tag{18}$$

In case of last activity sample in a video, we only remove one element ($p(A_q|A_{q-1})$). Finally, we compute an atypical score for each activity sample as, $\frac{\tilde{\mathcal{E}}_i}{N_t}$, where N_t denotes the number of tuples removed from the original sequence. Using the atypical scores, we can formulate our optimization problem to select the informative samples for manual labeling as discussed next.

D. Informative Sample Selection for an Activity Sequence

In this section, we will formulate an objective function from the atypical score of samples as discussed before. This objective function will be optimized to select the most informative samples. Let us consider that we have a batch of N unlabeled instances and we need to select the optimal instances for manual labeling. Let us define a vector $\boldsymbol{\mathcal{T}}_j = [\tilde{\mathcal{E}}_1 \quad \tilde{\mathcal{E}}_2 \quad \ldots]^T$, containing the atypical

Algorithm 1 Computation of Atypial Score and Uncertainty for Sample Selection

INPUTS. 1. Learned models from training data $\mathcal{L}$: Classification Model $\mathcal{M}$ and Transition Matrix T_s
2. Unlabeled Video Clips: $\mathcal{U}$
OUTPUTS. The vectors for atypical Score $\mathcal{T}$ and entropy $\boldsymbol{h}$ for the unlabeled data $\mathcal{U}$
Step 1: Compute Stationary Probability μ as shown in Eqn. 9 using T_s.
Step 2: Compute Entropy Rate $H(\mathcal{A})$ using μ and T_s as in Eqn. 15.
Step 3: Obtain an activity sequence from the predicted labels provided by $\mathcal{M}$ for a video clip in $\mathcal{U}$.
Step 4: Compute the probability of sequence $p(\mathcal{A}^P)$ using μ and T_s as in Eqn. 7
Step 5: Compute atypical score $\tilde{\mathcal{E}}_q$ using Eqn. 18 and entropy h_q associated with q^{th} sample.
Step 6: Calculate vectors $\mathcal{T}$ and $\boldsymbol{h}$ as discussed in Sec. III-D

score of each sample of the j^{th} video depending on the recognition task (e.g., activity recognition) as in Eqn 18.

Consider a vector $\boldsymbol{\mathcal{T}}$ which represents the atypical scores of the samples for all videos. We can write $\boldsymbol{\mathcal{T}}$ in terms of $\boldsymbol{\mathcal{T}}_j$ as follows.

$$\boldsymbol{\mathcal{T}} = [\boldsymbol{\mathcal{T}}_1 \quad \boldsymbol{\mathcal{T}}_2 \quad \dots]^T \tag{19}$$

We also incorporate the uncertainty of current baseline classifier on the unlabeled samples. We define a vector that denotes the entropy of all samples as $\boldsymbol{h} = [h_1 \quad h_2 \quad \dots]^T$, where $h_j = \mathbb{E}[-\log_2 p_j]$, and p_j is the probability mass function (p.m.f) which represents the distribution of prediction scores over the set of activity classes. This prediction score is generated by the current baseline classifier on the j^{th} unlabeled sample. We aim to choose a subset of the samples which are informative based on the two criterion, namely atypical score and the entropy of each sample obtained from the classifier. We can write the optimization function in vector form as follows,

$$\begin{aligned} \boldsymbol{y}^* &= \arg\max_{\boldsymbol{y}} \ \boldsymbol{y}^T(\boldsymbol{h} + \lambda\boldsymbol{\mathcal{T}}) \\ s.t. \quad & \boldsymbol{y} \in \{0,1\}^N \ , \quad \mathbf{y^T 1} \leq \eta \end{aligned} \tag{20}$$

Here, λ is a weighting factor. The term $\boldsymbol{y}^T\mathbf{1}$ represents the number of samples will be chosen which is bounded by η. Let us denote $\boldsymbol{f} = -(\boldsymbol{h} + \lambda\boldsymbol{\mathcal{T}})$. Maximization of the objective function in Eqn. 20 is the same as minimization of $\boldsymbol{y}^T\boldsymbol{f}$. It is a binary linear integer programming problem and can be solved by CPLEX [51]. Algorithm 1 shows the steps of our proposed method

Algorithm 2 Sample Selection for Active Learning with Continuous Data

INPUTS. 1. Learned models at Batch$_{k-1}$: Classification Model $\mathcal{M}_{k-1}$ and Transition Matrix T_s^{k-1}
2. Unlabeled Video Clips at Batch$_K$: $\mathcal{U}_k$
OUTPUTS. Learned Models after processing videos in Batch$_K$: $\mathcal{M}_k$ and T_s^k
Initialize: $\mathcal{L} = \{L_0\}$ (Initial Set of Data)
Step 1: Calculate vectors $\mathcal{T}$ and $\boldsymbol{h}$ using Algorithm 1
Step 2: Find optimal set of samples $\boldsymbol{y}_k^*$ using Eqn. 20 for Batch k
Step 3: $\mathcal{L} = \mathcal{L} \cup \boldsymbol{y}_k^*$
Step 4: Update models $\mathcal{M}_{k-1}$ and T_s^{k-1} with $\mathcal{L}$.

for selecting informative samples. Next, we show how the sample selection strategy can be used for active learning and anomaly detection as two applications for the experiments.

1) Active Learning: The sample selection strategy discussed above can be used in an active learning framework to update a classification model online. The adaptability of recognition models to the continuous data stream becomes important for long-term performance. Given a set of data at particular time, the proposed sample selection approach (discussed in Sec. III-D) can be utilized to select the most informative samples in order to update the model. After obtaining a set of samples $\boldsymbol{y}^*$ from Eqn. 20, we can ask a human to label these samples. With newly generated labeled data, the classification model $\mathcal{M}$, and the temporal relationship model need to be adjusted.

Update $\boldsymbol{\mathcal{M}}$. For classification task, we use softmax classifier to predict the labels. If the feature vector is $\mathcal{F}_k$ for k^{th} sample, then predicted probability for the j^{th} class can be written as, $P(l = j|\mathcal{F}_k) = \frac{e^{\mathcal{F}_k^T w_j}}{\sum_{k=1}^K e^{\mathcal{F}_k^T w_k}}$. Here, K is the number of classes, w_j represents the weights corresponding to class j. We optimize the cross entropy loss function to estimate the parameters as presented in [52]. For the current batch, we update the parameters with the newly labeled data samples.

Update Temporal Relationship Model. Let us consider a matrix Φ that represents the temporal statistics between activities. Φ will be updated based on the newly acquired labels. The updated statistics can be written as, $\Phi' \leftarrow \Phi + \tilde{\Phi}$, where $\tilde{\Phi}(.)$ represents the statistics with the newly labeled samples and Φ' is the updated statistics. With updated Φ, transition matrix T_s is modified.

2) Anomaly Detection: In anomaly detection, we consider an activity as abnormal if it is an outlier with respect to the learned model. Thus, any prior information

Algorithm 3 Algorithm of Proposed Method for Anomaly Detection

INPUTS. 1. Learned models with normal activities : Classification Model $\mathcal{M}$ and Transition Matrix T_s
2. Test Video Clip $\mathcal{V}_t$.
OUTPUTS. Set of binary labels $\mathcal{C}$ for anomaly and normal activities for $\mathcal{V}_t$.
Step 1: Compute atypical score $\tilde{\mathcal{E}}_j$ using Eqn. 18 and entropy h_j associated with j^{th} sample using Algorithm 1.
Step 2: Calculate irregularity score $\mathcal{D}_j$ using Eqn. 21.
Step 3: Assign class labels $\mathcal{C}$ for all the activities in $\mathcal{V}_t$ based on threshold τ as discussed in III-D2.

on anomalous activity at training time is unknown. We learn the recognition model $\mathcal{M}$ from the regular activities. The temporal relationship between the activity samples is also exploited during the learning process. We compute transition matrix T_s from this temporal relations. We calculate stationary distribution μ followed by entropy rate $H(\mathcal{X})$ using Eqns. 9 and 12.

Given a test video, $\mathcal{M}$ predicts the activity labels, from which a sequence is formed. We can compute atypical score $\tilde{\mathcal{E}}_j$ associated with j^{th} sample as discussed in Sec. III-C using Eqn. 18. We also compute the entropy $h_j = \mathbb{E}[-\log_2 p_j]$ from the distribution of confidence score provided by $\mathcal{M}$. We can now define irregularity score $\mathcal{D}_j$ which can be written as

$$\mathcal{D}_j = \tilde{\mathcal{E}}_j + \beta h_j. \tag{21}$$

β represents weighting factor. We also consider entropy along with the atypical score in order find an anomaly. Given an anomaly class, entropy should be high as it exhibits high uncertainty. All the steps are demonstrated in Algorithm 3. If $\mathcal{D}_j$ is larger than a threshold τ then it is considered as an abnormal class, or normal otherwise. The class of a sample C_j can be determined as follows.

$$C_j = \begin{cases} 1, & \text{if } \mathcal{D}_j > \tau \\ 0, & \text{otherwise} \end{cases} \tag{22}$$

Here, 1 represents abnormal activity and 0 denotes normal class. Next, we will demonstrate the experimental analysis of our proposed approach to sample selection and anomaly detection.

IV. Experiments

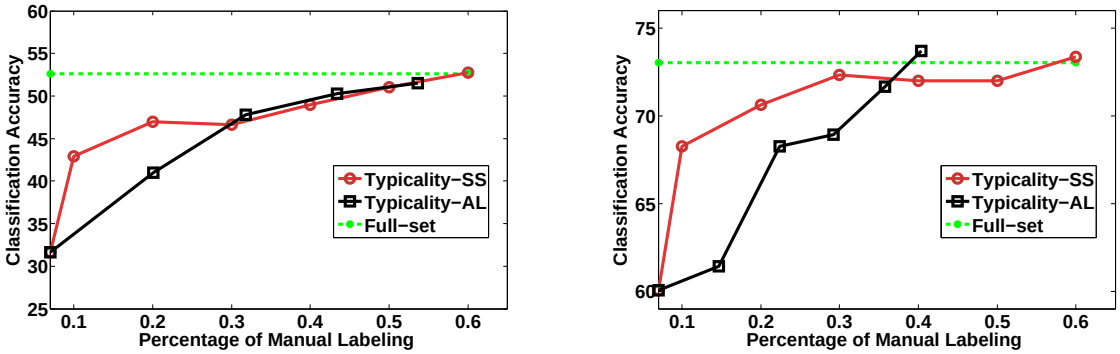
In this section, we evaluate our proposed method on two distinct applications such as informative sample selection for recognition model, and anomaly detection, for activity recognition task. We also compare our methods with state-of-the-art approaches on two challenging datasets.

Datasets. We demonstrate the performance of our proposed method on two video datasets. We evaluate our results on VIRAT [54] and MPII-Cooking [53] datasets for activity classification task. VIRAT is a video dataset which provides temporal relations between different activity samples. This dataset has 329 video clips consisting of 11 different activities [54]. MPII-Cooking dataset presents 65 cooking activities, e.g., *cut slices, pour, or spice* [53]. It has 44 videos in total. Since videos are usually long, we follow sliding window approach for cropping short video clips in order to create more video instances. We choose the video clip based on the number of activities. In this work, each video clip contains 8 activities with a stride of one activity for MPII-Cooking dataset.

Feature Extraction. For activity recognition model, we adopt the classification model described in [55]. We utilize the final layer of 3d convolutional neural network to extract features. Finally, we have 4096 dimensional c3d [55] feature for an activity sample (small clip of a video). These features are used to train softmax classifier for activity recognition.

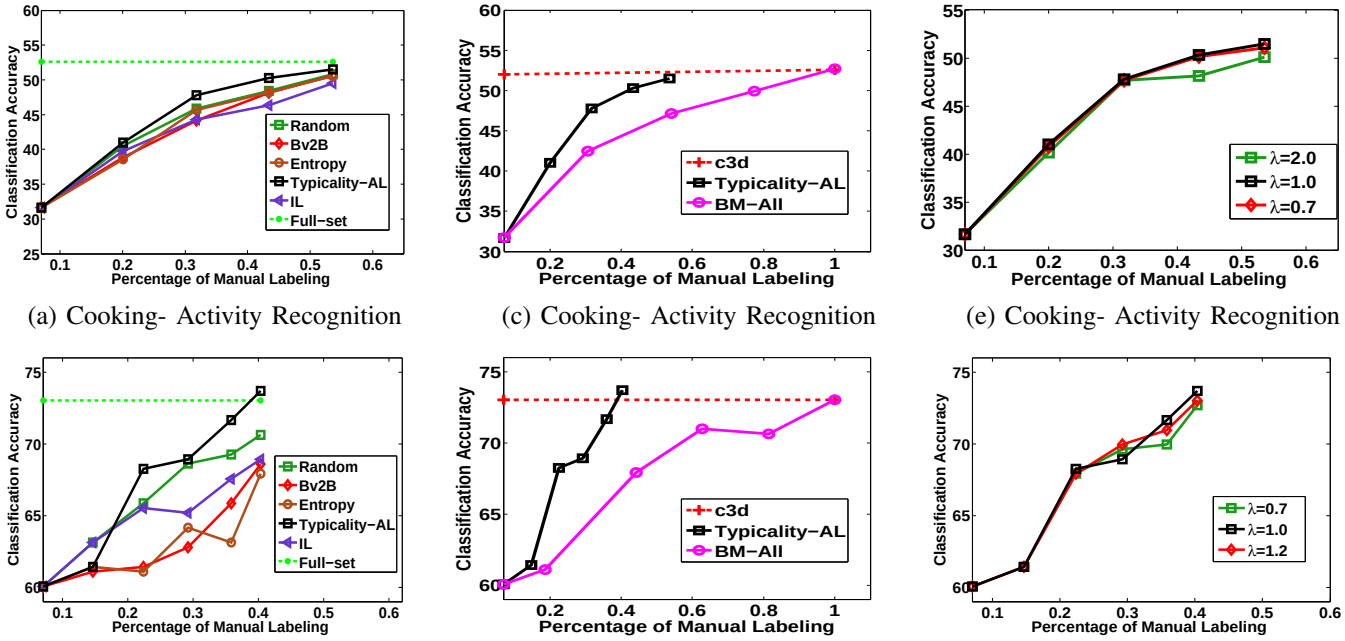
Evaluation Criterion. In order to evaluate active learning (AL) methods, we generate a plot of recognition accuracy vs percentage of manual labeling. We aim to achieve the same performance with less manual labeling effort. We utilize percentile (%) accuracy for activity recognition. For anomaly detection, we use ROC curve which measures the performance of binary classification task with varying threshold on prediction score. Finally, we calculate the area under the curve (AUC) to assess the performance. The value of AUC generally lies in between 0 and 1. We aim to achieve higher AUC value.

Experimental Setup. Our goal is to demonstrate two applications- (a) informative sample selection, and (b) anomaly detection, using proposed method discussed in Sec. III. In order to choose the most informative samples, we consider two scenarios- (1) sample selection from fixed data, (2) batch-mode active (online) learning. In first scenario, we fix the percentage of manual labeling from the whole training set and measure the performance on test set. In this setting, proposed framework inspects all the samples while selecting the informative samples. We learn the initial model from very few samples which are excluded from the training set. In batch-mode active learning, we consider same experimental setting as [7], where data samples (videos) are continuously coming in batches. We first divide the dataset into training and testing set. We create 5/6 batches from the training



(a) MPII-Cooking Dataset (b) VIRAT Dataset

Fig. 3: This figure illustrates the recognition performance of the proposed method for the tasks of informative sample selection and active learning, on (a) MPII-Cooking, and (b) VIRAT datasets.



(a) Cooking- Activity Recognition (c) Cooking- Activity Recognition (e) Cooking- Activity Recognition

(b) VIRAT- Activity Recognition (d) VIRAT- Activity Recognition (f) VIRAT- Activity Recognition

Fig. 4: The figure presents the performance of proposed active learning method for activity recognition task on two datasets - MPII-Cooking [53] (first row) and VIRAT [54] (second row) datasets. Plots (a,b) present the comparison against other state-of-the-art sample selection methods. Plots (c,d) demonstrate comparison with BM-All method. Plots (e,f) demonstrate the sensitivity analysis of our framework. Best viewable in color.

set. We evaluate the recognition performance on the test set after processing of each batch. Initial models (classification and temporal relations) are learned from the first batch of data. Typically, the first batch is smaller than other batches. Next, we apply sample selection strategy on next batches to choose the most informative samples. From the newly learned samples, models are updated. We also incorporate incremental learning to update the model as new classes can come in new batches. For anomaly detection, we learn the recognition and temporal relations from the normal activities. Now, given a test video, we compute irregularity score as discussed in Sec. III-D2, on which we determine whether an activity is anomalous or not.

State-of-the-art and Baseline Methods: In the experiment, we compare against different existing approaches and some baseline methods. These methods are as follows.

◇ **Typicality-SS:** Proposed approach applied to informative subset selection.

◇ **Typicality-AL:** Typicality based sample selection strategy for active (or online) learning task.

◇ **Bv2B:** Best vs Second Best active learning strategy [25].
◇ **IL:** Incremental learning approach presented in [56].
◇ **Full-set:** Entire training is used to obtain the accuracy from baseline classifiers.
◇ **BM-All:** All the samples in current batch are considered.
The baseline methods mentioned above are implemented on our training and testing set for fair comparison.

A. Informative Subset Selection from Fixed Data

In order to evaluate the performance of our sample selection strategy discussed in Sec. III-D, we vary the percentage of manual labeling from the training set, and measure the performance on test set. We keep the initial set fixed which is learned from very few samples. In this experimental setup, whole training set is observed in sample selection process. On the contrary, data are coming into batches for active learning. In our experiment, we choose 10% to 60% with 10% increment as the percentage of manual labeling, and compute the recognition accuracy for activity recognition. Fig. 3 illustrates the performance of our proposed method on sample selection. In this figure, we plot the classification accuracy of our proposed method with varying the percentage of manual labeling on two applications- sample selection and active learning. Typicality-SS and Typicality-AL represent the proposed approach for sample selection and active learning respectively. From this figure, we observe that the recognition performance of typicality-SS outperforms typicality-AL as the percentage of manual labeling decreases. The underlying reason is that typicality-SS considers the whole training set during the sample selection process unlike typicality-AL where active learner only utilize small portion of full dataset.

B. Performance of Batch-Mode Active Learning

We perform a various set of experiments to evaluate our proposed framework for online learning. We analyze the following experiments: 1. Comparison with existing active learning approaches, 2. Comparison against baseline methods, 3. Sensitivity analysis of the parameters, and 4. Time complexity of the proposed method, and 5. Performance with varying sequence length.

1) Comparison With Other Active Learning Methods: We compare our active learning (AL) approach with other state-of-the-art methods and baseline approaches as mentioned above. Figs. 4(a,b) show the recognition performance with respect to the percentage of manual labeling. We observe the performance on test set with updated recognition model after processing each batch of data. The straight line presented in the figures implies recognition accuracy of the model with 100% manual labeling (whole training set). We compare with some of the existing AL approaches such as Bv2B [25], random sample selection, Entropy [57] and IL [56]. For comparison, we first run our AL method to obtain the number of samples, which will be manually labeled. Then, we fix the number of samples for each batch and obtain the accuracy for other AL methods. In other words, different AL methods select the different subset of samples from each batch, where the size of subsets would be same. The performance will vary due to the selection of different subsets. For a fair comparison, we also keep same features and baseline classifiers for all the methods. From Figs. 4(a,b), we can see that the proposed framework *outperforms other AL methods to reduce the manual labeling cost by a large margin* in activity classification. Our method requires only **54%**, and **40%** of manual labeling to achieve the optimal recognition performance on VIRAT [54] and MPII-Cooking [53] datasets respectively as shown in Figs. 4(a,b). From Figs. 4(a,b), we can also see the performance gap between our method and other approaches. In MPII-Cooking [53] dataset, our approach outperforms Bv2B [25], random sample selection, Entropy [57] and IL [56] by **0.89%**, **0.67%**, **0.97%** and **2.00%** respectively with 54% manual labeling. Similarly, for VIRAT [54] dataset, proposed method surpasses Bv2B [25], random sample selection, Entropy [57] and IL [56] by **5.12%**, **3.07%**, **5.80%** and **4.78%** respectively with 40% annotation effort.

2) Comparison Against Other Baseline Methods: To evaluate proposed approach, we compare against BM-All method for activity classification. BM-All represents all the samples in a current batch, thus for N_b batches we have N_b accuracy values. Figs. 4(c,d) show the plots of our proposed model and BM-All method. BM-All helps us to understand the effectiveness of proposed

Dataset	Method	Accuracy (%)	Manual Labeling
MPII Cooking	GRP-Grassmann [21]	53.8%	100%
	C3D [55]	52.6%	100%
	C3D + SS (Ours)	51.49%	53.61%
VIRAT	Sparse AE [58]	54.2%	100%
	Joint Prediction [20]	71.8%	100%
	C3D [55]	73.03%	100%
	C3D + SS (Ours)	73.72%	40.36%

TABLE I: Comparison with other State-of-the-art methods on MPII-Cooking [53] and VIRAT [54] datasets. Here, the proposed model achieves similar or better performance as other methods with a fraction of manual labeling.

method in selecting the most informative samples. We aim to achieve similar performance with BM-All with less manual labeling effort. From the comparison of BM-All and proposed method, we can observe that a good recognition model can be learned from a small set of informative samples. Figs. 4(c,d) demonstrate that the proposed framework achieves similar or better performance with fewer informative samples when compared to BM-All method. In Fig. 4(d), we can also see that the proposed method outperforms the model with 100% labeling (red straight line). This also attests that informative (quality) data is often more useful than simply more data (quantity). Furthermore, we also compare against our method against other baseline methods with 100% manual labeling as shown in Table. I. From this table, we can see that the proposed sample selection method achieves better or similar performance with less manual labeling on MPII-Cooking [53] and VIRAT [54] datasets.

3) Sensitivity Analysis of the Parameters: In the proposed framework, we use the parameter λ as discussed in Sec. III-D. In order to understand the efficacy of typicality, we show different plots with varying λ in Figs. 4(e,f). We set the values of λ ranging from 0.7 to 2.0. We empirically choose these values to observe the change in plots. Figs. 4(e,f) illustrate the variation of performance due to change in hyperparameter λ. From the figures, we can see that the accuracy curve is stable with a little change in λ. The accuracy curve degrades with the smaller value of λ. If the value of λ equals to zero, the performance has been significantly dropped as presented as Entropy in figures Figs. 4(a,b) on two datasets.

4) Time Complexity. : The proposed method also reduces computation time to adapt the recognition model. Table. II shows the computational time on MPII-Cooking [53] and VIRAT [54] datasets. We compute the time to query the samples, and time to train recognition models for a dataset. We also compute the time to train a recognition model with all the samples in a batch (BM-All method). As we can see that total time to train activity model with all the samples is $3281.08s$ for MPII-Cooking [53], and $69.84s$ for VIRAT [54] dataset. On the other hand, the total time for querying and training with samples selected by our approach is $2498.32s$, and $62.75s$ for MPII-Cooking [53] and VIRAT [54] datasets respectively. In Table. II, we have also computed the execution time to query the samples by using Eqn. 20. The total query time for VIRAT and MPII-Cooking datasets is $0.1074s$ and $0.6113s$ respectively as provided in Table. II. We empirically observe that the query time at initial batches is longer than the query time at later batches. For instance, the query time in processing first batch is $0.048s$ and $0.563s$ for VIRAT and MPII-Cooking datasets respectively. From the Table. II, we can see that the proposed AL method helps to save a significant amount of computational time, especially in a big dataset.

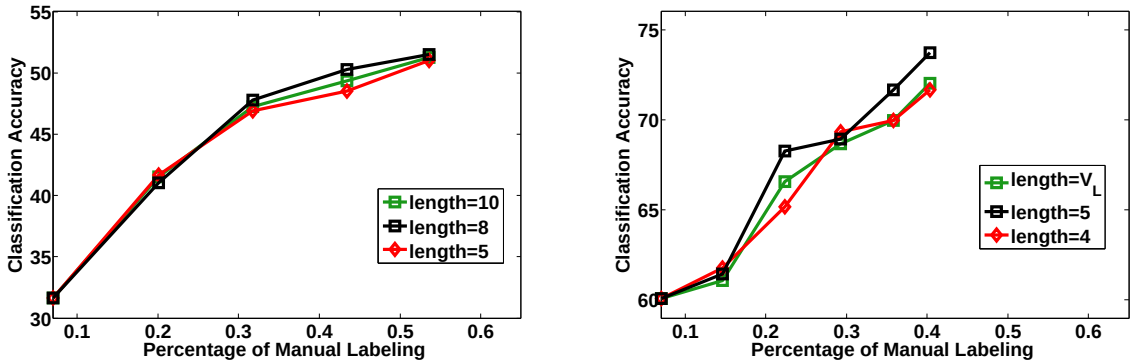
Method	Cooking [53] Time (s)	VIRAT [54] Time (s)
Query Time	$0.6113s$	$0.1074s$
Proposed Method	$62.75s$	$69.84s$
BM-All Method	$2498.32s$	$3281.08s$

TABLE II: Analysis of computation time on MPII-Cooking [53] and VIRAT [54] datasets. We can see from the table that our approach reduces computation time during training of recognition model. Query time represents the time to find a subset by using Eqn. 20.

5) Performance with Varying Sequence Length : We also set up an experiment in order to observe the effect of varying sequence length on recognition performance for active learning. We consider both MPII-Cooking and VIRAT datasets to run this experiment. Sequence length represents the number of activities in a video clip. In order to prepare the data, we extract video clip with varying sequence length from the original video by following sliding window. For MPII-Cooking dataset, we vary the sequence length to $10, 8$ and 5. We consider the whole video (length=V_L) and two different lengths (4 and 5) for VIRAT dataset. The performance of proposed method with varying sequence length on MPII-Cooking and VIRAT datasets is demonstrated in Fig. 5(a) and 5(b) respectively. From the figures, higher performance is observed for the moderate sequence length (V_L) (e.g., $V_L = 8$ and $V_L = 5$ for MPII-Cooking and VIRAT datasets respectively), when compared to the small sequence length ($V_L = 5$ for MPII-Cooking and $V_L = 4$ for VIRAT). However, we do not observe the best performance for a longer sequence. The underlying reason is that the rate of misclassified activity samples is higher in a long sequence which restricts the typicality model to capture good temporal relationship between the samples. In order to obtain similar performance as the moderate sequence, the amount of manual labeling will be higher for a long sequence.

C. Anomaly Detection

In this section, we will show how atypical score (discussed in Sec. III-C) can be utilized to detect anomaly activities. In this paper, we perform N_a-fold cross validation in order to evaluate the performance of anomaly detection task. Here, N_a is the number of activity categories. In each fold, one class is chosen



(a) MPII-Cooking Dataset (b) VIRAT Dataset

Fig. 5: This figure illustrates the recognition performance with varying sequence length on (a) MPII-Cooking, and (b) VIRAT datasets.

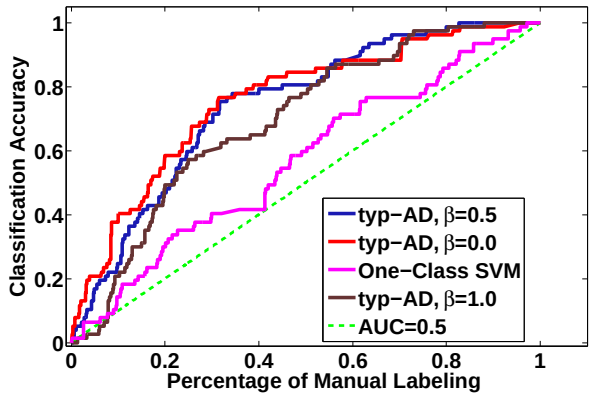


Fig. 6: The figure shows ROC plots for anomaly detection on VIRAT [54] dataset. Different colors represent baseline methods. Best viewable in color.

as abnormal and rest are utilized as normal activities. The process of choosing abnormal class iterates over all the classes, which gives us a larger evaluation set. We learn the temporal relations from normal activity classes. To train the recognition model, we train multi-class softmax classifier with normal activities. We exclude the abnormal class during the learning process. Given a test video, recognition model provides a probability distribution over the classes. We compute entropy from this distribution. If the activity class belongs to normal activities, recognition model shows low uncertainty. For abnormal class, the uncertainty goes high. We also calculate the atypical score for the activity samples in test video. After computing the uncertainty and atypical score, we calculate the irregularity score using Eqn. 21. Based on this irregularity score, we determine whether an activity is abnormal or not.

In order to evaluate our framework, we plot ROC curve by varying the threshold on irregularity. Fig. 6 shows ROC plots for different methods on VIRAT [54] dataset. VIRAT dataset provides the ground-truth label (the category of an activity) for each sample. Our proposed method relies on the activity category for modeling a sequence of activities and hence we choose this dataset. We consider One-class SVM [59] as the baseline method to detect anomalous activity. For convenience, we refer our proposed method as 'typ-AD' (i.e, typicality for Anomaly Detection) in Fig. 6. In typ-AD, we change the value of β as discussed in Sec. III-D2 ranging from 0 to 1.0 to observe the effect on uncertainty score. From the figure, we can see that the proposed method with only atypical score ($\beta = 0$) outperforms all other methods. As the value of β increases, we put more weights on entropy h_j as shown in Eqn. 21. We can see from the Fig. 6 that the performance of anomaly detection is improved as the value of β decreases.

We also provide area under the curve (AUC), which is computed from ROC curves as shown in Fig. 6 to measure the performance of anomaly detection. Table III illustrates the value of AUC for different methods. As we change the value of β, we observe different performance. With values of $\beta = 0.5$ and $\beta = 1.0$, we obtain AUC of 0.75 and 0.70 respectively. We observe the best performance with $\beta = 0$, which achieves 0.76 in AUC value. In the anomaly detection problem, the behavior of anomaly samples is absolutely unknown to the recognition model as the anomaly samples are chosen from a new category that is excluded from the training set. In ideal condition, the entropy explained in Sec. III-D2 should be high for an anomalous sample given a good recognition model. However, we empirically observe that the recognition model provides a wrong label with a high confidence score for a test sample in many cases. As a result, entropy goes down, and the entropy has less

Method	AUC Score
typ-AD with $\beta = 1.0$	0.70
typ-AD with $\beta = 0.5$	0.75
One-Class SVM [59]	0.57
Context-Aware Model [11]	0.685
typ-AD with $\beta = 0$	0.76

TABLE III: The table illustrates the performance of anomaly detection in terms of AUC score on VIRAT [54] dataset.

impact on determining the irregularity value as presented in Eqn. 21. In contrast, an anomalous activity violates the properties of typical set by impairing the temporal behavior pattern of the activity samples. So, the model with $\beta = 0$ shows superior performance in anomaly detection. We also compare against other methods such as one-class SVM [59] and Context-Aware Model [11]. The AUC values for One-class SVM [59] and Context-Aware Model [11] are 0.57 and 0.685 respectively.

1) Limitations of the Proposed Approach in Anomaly Detection: In this paper, we exploit the notion of typicality for the task of active learning and show the preliminary work in detecting contextual anomalies. The typicality model as presented in Sec. III-D2 exploits the temporal link between two consecutive activities. Thus, the proposed approach mainly focuses on detecting the anomalous samples which are temporally inconsistent. In order to learn these temporal relations, an activity category or label is explicitly utilized. However, most of the conventional anomaly datasets, e.g., UCSD Pedestrian [60], CUHK Avenue [13], Subway [61], do not provide the category of an activity for a sample. Instead, these datasets provide a binary label for the normal and abnormal class. Since they do not have the labels of the activities, we cannot evaluate our anomaly detection results on these datasets without extensively labeling all the activities first. Thus, we demonstrate our results on VIRAT dataset which provides the ground-truth labels for each activity sample thus allowing us to evaluate the performance of our method.

An activity sample is identified as anomaly based on the irregularity score presented in Eqn. 21. We learn the transition matrix from the temporal interactions between the activity samples as shown in Eqn. 14. In Eqn. 14, ϕ_k^i represents the number of appearing activity class a_k with previous activity a_i. Please note that an anomalous sample belongs to a new class which is unknown to the recognition model. In most of the cases, the recognition model predicts a wrong label with which the previous activity and next activity do not show strong temporal correlation for an anomalous sample (ϕ_k^i is low). As a result, the irregularity score in Eqn. 21 becomes high and the sample is detected as an anomaly. However, in very few cases, the recognition model provides a label which might show high temporal correlation with last activity even though the classifier is not confident enough (ϕ_k^i is high). In such cases, our proposed typicality based model fails to detect an anomaly. In the future, we intend to present an anomaly detection model for identifying collective anomalies and taking into account spatial inter-relationships in addition to temporal ones.

V. Conclusions

In this paper, we presented a subset selection method by exploiting information-theoretic 'typical set' to adaptively learn the recognition models. We show that typicality is a powerful tool which has been successfully used in data processing and can also be utilized in informative subset selection problem for visual recognition tasks. Our method is applied to various applications including sample selection and anomaly detection. The notion of typicality is used for a sequence of activities that can be represented as a Markov chain. Our approach significantly reduces the human load in labeling samples for visual recognition tasks. We demonstrate that our method achieves better or similar performance with only a small subset of the full training set compared with a model using full training set. Our model also shows good performance in anomaly detection in a video. As a future direction, we will study how typicality can be utilized to transfer knowledge from one domain where data is available to another where there is limited labeled data.

Acknowledgement. This work was partially funded by NSF grant 1316934 and 1544969.

References

[1] A. Lapedriza, H. Pirsiavash, Z. Bylinskii, and A. Torralba, "Are all training examples equally valuable?" *arXiv preprint arXiv:1311.6510*, 2013. 1

[2] B. Settles, "Active learning," *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 6, no. 1, pp. 1–114, 2012. 1, 3

[3] X. Li and Y. Guo, "Multi-level adaptive active learning for scene classification," in *ECCV*, 2014. 1, 3

[4] X. Li et al., "Adaptive active learning for image classification," in *CVPR*, 2013. 1, 3

[5] M. Hasan and A. K. Roy-Chowdhury, "Context aware active learning of activity recognition models," in *ICCV*, 2015. 1, 2, 3

[6] J. H. Bappy, S. Paul, and A. Roy-Chowdhury, "Online adaptation for joint scene and object classification," in *ECCV*, 2016. 1, 2, 3

[7] J. H. Bappy, S. Paul, E. Tuncel, and A. K. Roy-Chowdhury, "The impact of typicality for informative representative selection," *CVPR*, 2017. 1, 2, 3, 4, 8

[8] S. Paul, J. H. Bappy, and A. Roy-Chowdhury, "Non-uniform subset selection for active learning in structured data," in *CVPR*, 2017. 1, 2, 3

[9] O. P. Popoola and K. Wang, "Video-based abnormal human behavior recognitiona review," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 865–878, 2012. 1, 4

[10] K. Cheng, Y. Chen, and W. Fang, "Video anomaly detection and localization using hierarchical feature representation and gaussian process regression," in *CVPR*, 2015. 1, 4

[11] Y. Zhu, N. M. Nayak, and A. K. Roy-Chowdhury, "Context-aware activity recognition and anomaly detection in video," *IEEE Journal of Selected Topics in Signal Processing*, vol. 7, no. 1, pp. 91–101, 2013. 1, 3, 4, 13

[12] V. Saligrama and Z. Chen, "Video anomaly detection based on local statistical aggregates," in *CVPR*, 2012. 1

[13] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in matlab," in *ICCV*, 2013. 1, 13

[14] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *CVPR*, 2011. 1

[15] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *CVPR*, 2016. 1, 4

[16] T. M. Cover and J. A. Thomas, *Elements of information theory*. John Wiley & Sons, 2012. 2, 4

[17] W. Choi, K. Shahid, and S. Savarese, "Learning context for collective activity recognition," in *CVPR*, 2011. 3

[18] F. J. Ordóñez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016. 3

[19] M. Zeng, L. T. Nguyen, B. Yu, O. J. Mengshoel, J. Zhu, P. Wu, and J. Zhang, "Convolutional neural networks for human activity recognition using mobile sensors," in *International Conference on Mobile Computing, Applications and Services (MobiCASE)*, 2014, pp. 197–205. 3

[20] T. Mahmud, M. Hasan, and A. K. Roy-Chowdhury, "Joint prediction of activity labels and starting times in untrimmed videos," in *ICCV*, 2017. 3, 10

[21] A. Cherian, S. Sra, S. Gould, and R. Hartley, "Non-linear temporal subspace representations for activity recognition," in *CVPR*, 2018. 3, 10

[22] A. Cherian, B. Fernando, M. Harandi, and S. Gould, "Generalized rank pooling for activity recognition," *arXiv preprint arXiv*, vol. 170402112, 2017. 3

[23] C. Käding, A. Freytag, E. Rodner, A. Perino, and J. Denzler, "Large-scale active learning with approximations of expected model output changes," in *German Conference on Pattern Recognition*, 2016. 3

[24] B. Settles, "Active learning literature survey," *University of Wisconsin, Madison*, vol. 52, no. 55-66, 2010. 3

[25] X. Li, R. Guo, and J. Cheng, "Incorporating incremental and active learning for scene classification," in *ICMLA*, 2012. 3, 10

[26] E. Elhamifar, G. Sapiro, A. Yang, and S. Shankar Sasrty, "A convex optimization framework for active learning," in *ICCV*, 2013. 3

[27] D. Xu, R. Song, X. Wu, N. Li, W. Feng, and H. Qian, "Video anomaly detection based on a hierarchical activity discovery within spatio-temporal contexts," *Neurocomputing*, vol. 143, pp. 144–152, 2014. 3

[28] Y. Zhang, H. Lu, L. Zhang, X. Ruan, and S. Sakai, "Video anomaly detection based on locality sensitive hashing filters," *Pattern Recognition*, vol. 59, pp. 302–311, 2016. 4

[29] J. Kim and K. Grauman, "Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates," in *CVPR*, 2009. 4

[30] T. Xiao, C. Zhang, and H. Zha, "Learning to detect anomalies in surveillance video," *IEEE Signal Processing Letters*, vol. 22, no. 9, pp. 1477–1481, 2015. 4

[31] M. J. Roshtkhari and M. D. Levine, "Online dominant and anomalous behavior detection in videos," in *CVPR*, 2013. 4

[32] M. Sabokrou, M. Fathy, M. Hoseini, and R. Klette, "Real-time anomaly detection and localization in crowded scenes," in *CVPR*, 2015. 4

[33] R. Hinami, T. Mei, and S. Satoh, "Joint detection and recounting of abnormal events by learning deep generic knowledge." in *ICCV*, 2017. 4

[34] R. Tudor Ionescu, S. Smeureanu, B. Alexe, and M. Popescu, "Unmasking the abnormal events in video," in *ICCV*, 2017. 4

[35] D. J. MacKay and D. J. Mac Kay, *Information theory, inference and learning algorithms*. Cambridge university press, 2003. 4

[36] S. H. Lim, C.-Y. Wang, and M. Gastpar, "Information-theoretic caching: The multi-user case," *IEEE Transactions on Information Theory*, vol. 63, no. 11, pp. 7018–7037, 2017. 4

[37] P. Noorzad, M. Effros, and M. Langberg, "The unbounded benefit of encoder cooperation for the k-user mac," *IEEE Transactions on Information Theory*, vol. 64, no. 5, pp. 3655–3678, 2018. 4

[38] Y. Suhov and I. Stuhl, "On principles of large deviation and selected data compression," *arXiv preprint arXiv:1604.06971*, 2016. 4

[39] A. Høst-Madsen, E. Sabeti, and C. Walton, "Data discovery and anomaly detection using atypicality: Theory," *arXiv preprint arXiv:1709.03189*, 2017. 4

[40] J. T. Maxfield, W. D. Stalder, and G. J. Zelinsky, "Effects of target typicality on categorical search," *Journal of vision*, vol. 14, no. 12, pp. 1–1, 2014. 4

[41] B. Saleh, A. Elgammal, and J. Feldman, "The role of typicality in object classification: Improving the generalization capacity of convolutional neural networks," *IJCAI*, 2016. 4

[42] J. Vogel and B. Schiele, "A semantic typicality measure for natural scene categorization," in *Joint Pattern Recognition Symposium*. Springer, 2004, pp. 195–203. 4

[43] T. M. Cover and J. A. Thomas, "Elements of information theory 2nd edition," 2006. 5

[44] P. H. Algoet and T. M. Cover, "A sandwich proof of the shannon-mcmillan-breiman theorem," *The annals of probability*, pp. 899–909, 1988. 5

[45] "Stationary distributions of markov chains," https://brilliant.org/wiki/stationary-distributions/, retrieved: May 10, 2018. 5

[46] M. Liu, O. Tuzel, S. Ramalingam, and R. Chellappa, "Entropy rate superpixel segmentation," in *CVPR*, 2011. 5

[47] N. Vaswani, A. K. Roy-Chowdhury, and R. Chellappa, "" shape activity": a continuous-state hmm for moving/deforming shapes with application to abnormal activity detection," *IEEE Transactions on Image Processing*, vol. 14, no. 10, pp. 1603–1616, 2005. 6

[48] V. Kellokumpu, M. Pietikäinen, and J. Heikkilä, "Human activity recognition using sequences of postures." in *MVA*, 2005, pp. 570–573. 6

[49] L. Liao, D. Fox, and H. Kautz, "Location-based activity recognition," in *NIPS*, 2006, pp. 787–794. 6

[50] N. M. Nayak, Y. Zhu, and A. K. R. Chowdhury, "Hierarchical graphical models for simultaneous tracking and recognition in wide-area scenes," *IEEE Transactions on Image Processing*, vol. 24, no. 7, pp. 2025–2036, 2015. 6

[51] I. I. CPLEX, "V12. 1: Users manual for cplex," *International Business Machines Corporation*, vol. 46, no. 53, p. 157, 2009. 7

[52] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Annals of operations research*, 2005. 7

[53] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A database for fine grained activity detection of cooking activities," in *CVPR*, 2012. 8, 9, 10, 11

[54] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. Aggarwal, H. Lee, L. Davis *et al.*, "A large-scale benchmark dataset for event recognition in surveillance video," in *CVPR*, 2011. 8, 9, 10, 11, 12, 13

[55] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatio-temporal features with 3d convolutional networks," in *ICCV*, 2015. 8, 10

[56] M. Hasan and A. Roy-Chowdhury, "Incremental activity modeling and recognition in streaming videos," in *CVPR*, 2014. 10

[57] G. Druck, B. Settles, and A. McCallum, "Active learning by labeling features," in *EMNLP*, 2009. 10

[58] M. Hasan and A. K. Roy-Chowdhury, "Continuous learning of human activity models using deep nets," in *ECCV*, 2014. 10

[59] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural computation*, vol. 13, no. 7, pp. 1443–1471, 2001. 12, 13

[60] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *CVPR*, 2010. 13

[61] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE transactions on pattern analysis and machine intelligence*, vol. 30, no. 3, pp. 555–560, 2008. 13

Jawadul H. Bappy received the B.S. degree in Electrical and Electronic Engineering from the Bangladesh University of Engineering and Technology, Dhaka in 2012. He received his Ph.D. in Electrical and Computer Engineering from the University of California, Riverside in 2018. He is currently working as a scientist at JD.Com in Mountain View, CA. His main research interest includes computer vision, media forensics, deep generative models, and advanced machine learning techniques for real-life applications.

Sujoy Paul is currently pursuing his Ph.D. degree in the Department of Electrical and Computer Engineering at the University of California, Riverside. He received his undergraduate degree in Electrical and Telecommunication Engineering from Jadavpur University. His broad research interest includes Computer Vision and Machine Learning with specialized interests on learning with limited supervision for action recognition, object detection, image and video captioning, tracking.

Ertem Tuncel [S'99-M04-SM18] received the B.S. degree in electrical and electronics engineering from the Middle East Technical University, Ankara, Turkey in 1995, the M.S. degree in electrical and electronics engineering from Bilkent University, Ankara, Turkey in 1997, and the Ph.D. degree in electrical and computer engineering from University of California, Santa Barbara, in 2002. In 2003, he joined the Department of Electrical and Computer Engineering, University of California, Riverside, where he is now a Professor. He also serves as the Associate Dean of the Graduate Division at the same university. Dr. Tuncel received the National Science Foundation CAREER Award in 2007. He was an Associate Editor of the IEEE TRANSACTIONS ON INFORMATION THEORY between May 2014 and June 2017. His main research interests are information theoretic analyses of multi-user networks, joint source-channel coding, zero-delay communication, energy-distortion tradeoffs, and content-based retrieval in high-dimensional databases.

Amit K. Roy-Chowdhury received the Bachelors degree in Electrical Engineering from Jadavpur University, Calcutta, India, the Masters degree in Systems Science and Automation from the Indian Institute of Science, Bangalore, India, and the Ph.D. degree in Electrical and Computer Engineering from the University of Maryland, College Park. He is a Professor of Electrical and Computer Engineering and a Cooperating Faculty in the Department of Computer Science and Engineering, University of California, Riverside. His broad research interests include computer vision, image processing, and vision-based statistical learning, with applications in cyber-physical, autonomous and intelligent systems. He is a coauthor of two books: Camera Networks: The Acquisition and Analysis of Videos over Wide Areas, and Recognition of Humans and Their Activities Using Video. He is the editor of the book Distributed Video Sensor Networks. He has been on the organizing and program committees of multiple computer vision and image processing conferences and is serving on the editorial boards of multiple journals. He is a Fellow of the IEEE and IAPR.