The Impact of Typicality for Informative Representative Selection

Jawadul H. Bappy, Sujoy Paul, Ertem Tuncel and Amit K. Roy-Chowdhury
Department of ECE, University of California, Riverside, CA 92521, USA
{mbappy,supaul,ertem,amitrc}@ece.ucr.edu

Abstract

In computer vision, selection of the most informative samples from a huge pool of training data in order to learn a good recognition model is an active research problem. Furthermore, it is also useful to reduce the annotation cost, as it is time consuming to annotate unlabeled samples. In this paper, motivated by the theories in data compression, we propose a novel sample selection strategy which exploits the concept of typicality from the domain of information theory. Typicality is a simple and powerful technique which can be applied to compress the training data to learn a good classification model. In this work, typicality is used to identify a subset of the most informative samples for labeling, which is then used to update the model using active learning. The proposed model can take advantage of the inter-relationships between data samples. Our approach leads to a significant reduction of manual labeling cost while achieving similar or better recognition performance compared to a model trained with entire training set. This is demonstrated through rigorous experimentation on five datasets.

1. Introduction

One of the challenges in visual recognition tasks is to learn a good classification model from a set of labeled examples. Today we live in a time where we have instant access to huge amount of visual data from online sources such as Google, Yahoo, Bing and Youtube. It becomes infeasible to label all the unlabeled samples as it is very costly and time consuming. Moreover, it is not always true that more labeled data can help a classifier to learn better; in fact, it may as well confuse the classifier [25]. Also, the adaptability of recognition models is unavoidable in order to achieve good classification performance that is robust to concept drift. As a result, selection of the most informative samples [41] becomes critical and has drawn significant recent attention from the vision community in order to train recognition models [40, 29]. Motivated by this, the goal of this paper is *to obtain a subset of few informative samples from the huge corpus of available unlabeled data to learn a good recognition model*.

In order to identify the informative samples, most active learning based query selection techniques choose the samples about which the classifier is most uncertain [40]. Recent advances in active learning exploit the inter-relationships between samples in order to reduce the number of labeled samples to train the models [27, 32], with applications in several recognition tasks such as activity recognition [18], and scene and object classification [2]. The utilization of context in active learning is sometimes referred as *context-aware active learning*. Most of the context-aware recognition tasks involve graphical models [36] to correlate between the samples. In order to measure uncertainty on a graph [48], we require node entropy as well as mutual information. It is shown in [48] that node entropy is calculated from node potential, and mutual information is computed from both node and edge potential. In recognition tasks, node potentials are usually designed from the classification score of the samples. Thus, a sample might not be selected if the classification score is high enough for the wrong class. Furthermore, it becomes computationally expensive or intractable to compute the mutual information when the number of random variables increases, and hence the above-mentioned methods need to make simplifying assumptions.

In this paper, we explore whether information theoretic ideas that have been very successfully applied in data compression can be used to identify the most informative samples to build a recognition model. We leverage upon the concept of typicality for this purpose. Typicality allows representation of any sequence using entropy as a measure of information. The concept of typical set is developed based on the intuitive notion that not all the messages are equally important, i.e., some messages carry more information than others. According to the theory, there is a set of messages for which the total probability of any of its members occurring is close to one, which is referred as typical set of messages. By analogy, in computer vision perspective, we are convinced that not all the samples are equally important to learn a recognition model. Thus, we ask how can we exploit this approach to select the most informative samples, which will be manually labeled, and classifiers designed on this subset can then be applied to the entire dataset. Although, the term 'typicality' is mentioned in some computer vision

papers for several tasks such as category search [34], object recognition [39], and scene classification [45], they do not exploit the notion of information-theoretic *typical set* as we aim to in this work.

In order to exploit the typicality in an image, we use the labels provided by the detectors as the elements to form a sequence. Thus, if a sequence deviates from typical set for an element of that sequence, the element has good chance to be selected for labeling. The major advantages of using typicality are the following.

(1) Typicality identifies a small subset of samples, which represents common characteristics of a class.

(2) Previously, in computer vision, one of the effective way to incorporate context into a recognition scheme was through a graphical model, where node potential is learned from classification score and edge potential is learned from contextual relations between the samples. In this paper, we show that typicality can also be used to link between recognition and context models.

(3) Typicality is computationally efficient. We can capture higher order relations among the elements of a sequence by exploiting typicality. Thus, we can apply this method when inter-relationship between data points is known. For example, in joint scene-object classification, typicality links all the detected objects with a scene, e.g. '*bed, lamp, painting, curtain* with scene *bedroom*'. On the other hand, graph based models consider the pair-wise constraint such as '*bed* in a *bedroom*', '*lamp* in a *bedroom*' to interlink between scene and objects.

(4) We can apply this technique in feature space as well, to find the informative samples by identifying the typical feature for a class.

Framework Overview: The flow of the proposed framework is shown in Fig. 1. The method starts with a small subset of labeled samples to build the initial classification model. We also learn the co-occurrence statistics of the samples when available. Our goal is to update these models with the manually labeled samples that will be selected by our proposed approach, leading to an active learning framework. We first classify samples using the current models for a batch of incoming unlabeled data. Then, we compute the entropy from the distribution of classification scores to obtain the uncertainty of the labels being predicted. In order to exploit typicality, we need to obtain a sequence and a distribution from which a sequence is drawn. We refer to this distribution as 'typical model'. We learn the typical model in two steps: (1) from the feature values, and (2) from the contextual relationship between samples when available. We obtain a sequence and distribution in feature space to find the atypical score (please see details in Sec. 3) associated with a sample. Similarly, for the latter case, we generate a sequence from the labels of the samples provided by the classifier, and learn the probability mass function (pmf) from the co-occurrence statistics of the samples. Finally, we formulate an optimization function in order to select the most informative set of samples based on entropy and typicality. The labels obtained in this process are used to update the classification model as well as the contextual relationships.

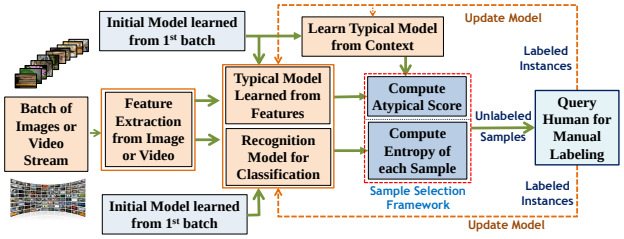


Figure 1: Overview of proposed framework to choose the most informative samples to train the recognition model.

Contributions: Our *major contributions* are as follows.

• We propose a general active learning framework by introducing 'typicality' concept from information theory. To the best of our knowledge, any previous work that uses *typicality for active learning* is unknown.

• We explicitly show how typicality can be used to find out contextual irregularity. We also determine the typical feature for a class which is very useful in recognition.

• Unlike most of the context-aware active learning approaches, we do not require a graph to inter-relate the samples, which makes our active learning method faster.

We demonstrate our experimental results on two scenarios- (1) multi-task classification such as scene-object and activity-object, (2) single-task classification like scene or object recognition. Our framework outperforms state-of-the-art methods significantly in reducing the manual labeling cost while achieving same recognition performance.

1.1. Related Works

We briefly review the related works in visual recognition, and then provide an overview of sample selection strategy.

Classification in Computer Vision. The proposed framework applies to several recognition tasks, such as scene, object and activity classification. A review paper in [43] discusses some of the common features such as color, texture and SIFT descriptor, which are used in image classification. In [37], the paper surveys state-of-the-art feature based activity recognition. Recent advances in computer vision use context model [3] on top of recognition model in order to achieve higher accuracy. The use of context model has been applied in several applications such as object recognition [35, 47], scene classification [49, 47, 1] and activity recognition [18]. Another promising approach in recognition tasks is to exploit deep learning. Deep learning based methods have achieved superior performance in recognition tasks such as scene classification [50], object detection [15, 20, 14] and activity recognition [18].

Sample Selection Methods. Active learning has been widely used to reduce the effort of manual labeling in different computer vision tasks including scene classifica-

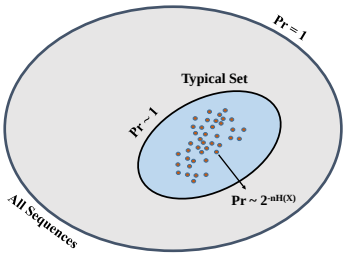


Figure 2: The figure illustrates the idea of typical set of sequences used in information theory.

tion [30, 28, 10], video segmentation [12], object detection [44, 23, 7], activity recognition [18] and tracking [46]. In active learning, some state-of-the-art approaches consider expected change in gradients [41], information gain [30], and expected prediction loss [29] to obtain the samples for querying. Some of the common techniques to measure uncertainty for selecting the informative samples are presented in [40, 29]. In [28], the authors incorporated two strategies - best vs. second best and K-centroid to select the informative subset. An active learning framework for object categories was proposed in [22] which considers the case where the labeler itself is uncertain about labeling an image.

The afore-mentioned approaches consider the individual samples to be independent. In [21], social relations were exploited for active learning of a text classification model in micro-blogging data. Spatial information was exploited in [27] to classify hyper-spectral images in an active learning framework. In [32] an active learning framework was proposed, which exploits the similarity between data points as relations between them in the feature space. In [18], contextual relationships between activities was exploited in an active learning framework for activity recognition. In [30], the authors presented a hierarchical active learning framework for scene classification. A recent paper [2] proposed a graph based active learning framework for joint scene-object recognition by exploiting contextual relationships.

2. Typicality in Information Theory

In information theory, typical set [4] is a collection of sequences, the total probability of whose occurrence is close to one as shown in Fig. 2. There are two types of typical sequences, which are generally used, namely, weak typicality and strong typicality. In this problem, we focus on weak typicality to design our active learning framework.

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

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

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:

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

We can derive a number of properties from this definition as $n \rightarrow \infty$ [4].

Property 1. The probability of any sequence in the typical set would be in the range below:

$$2^{-n[H(X)+\epsilon]} \leq P_{X^n}(\boldsymbol{x}^n) \leq 2^{-n[H(X)-\epsilon]} \quad (4)$$

It comes directly from the definition shown in Eqn. 4. If the value of ϵ becomes zero, then the probability of all sequences belonging to typical set is equal. In this paper, we show how to generate a sequence from the samples for a recognition task. In active learning application, we focus on the sequence which has probability outside the range of typical set as shown in Eqn. 4.

Property 2. The typical set has size of approximately $2^{nH(X)}$ sequences.

Property 3. The probability of a sequence drawn from typical set $\mathcal{A}_\epsilon$:

$$P[X \in \mathcal{A}_\epsilon^{(n)}] \geq 1 - \epsilon \quad (5)$$

For smaller ϵ , this probability reaches close to 1.

Property 4. More likely sequence might not be a member of typical set. Let us consider a vision problem, classification of joint scene and objects. Suppose that we have an i.i.d distribution, and we denote a random variable O. Here, $O \in \{$ bed(o_1), sofa$(o_2)\}$ and S represents bedroom. We have a distribution, $P(O = o_1|S) = 0.9$, $P(O = o_2|S) = 0.1$. So, the sequence $(o_1, o_1, ...o_1)$ is highly likely. However, it is not a typical set because its average probability is not close to the entropy of $P(O|S)$. This example signifies that even though *'bed'* has high co-occurrence probability with *'bedroom'*, typicality also tells us that the *'sofa'* detector is not working properly.

3. Typicality for Visual Recognition Tasks

In this section, we show how typicality can be used in feature space, as well as to model visual context, e.g., inter-relationships between objects in a scene.

Typicality with Contextual Relationships. Classification tasks, like scene-objects and activity-objects, generally share contextual relationship between the data points. We use typicality as a tool to capture these contextual relationships. For example, in joint scene-object classification, typicality encodes what are the typical objects that appear in a scene. We present below how contextual relations can be incorporated to compute the typical score.

In typicality, the sequences are generated from a distribution $P_{X^n}(.)$. We model this distribution as the co-occurrence frequency of one type of instance with another

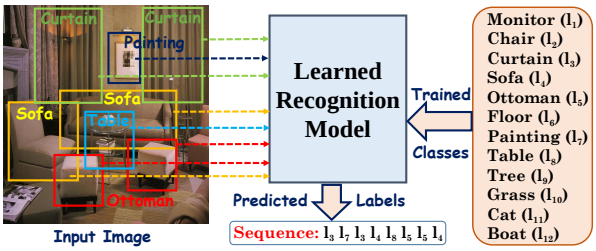


Figure 3: The figure shows how a sequence is generated from a recognition model in the context of object detection. The labels provided by the detectors (object recognition model) are used to represent a sequence.

type of instance, e.g., the co-occurrence of an object type given a scene type. In joint scene-object scenario, as multiple objects appear in a scene, we can find a co-occurrence distribution given a scene. Similarly, for joint activity and object classification, we consider a sequence for objects conditioned upon activity class. We assume that object detectors are running independently.

Let us consider two different classification tasks $\mathcal{U}$ and $\mathcal{V}$. We also assume that instances belonging to task $\mathcal{U}$ might co-occur multiple times with instances belonging to task $\mathcal{V}$. For instance, multiple objects can appear in a scene or activity, so object recognition would be $\mathcal{U}$ and scene or activity would be $\mathcal{V}$ in this case. Let us denote the number of classes for $\mathcal{U}$ and $\mathcal{V}$ as M and N. The co-occurrence frequency of the classes in $\mathcal{U}$ given the i^{th} class in $\mathcal{V}$ can be denoted by $\Phi(u|v_i) = [\phi_1^i, \phi_2^i, \ldots, \phi_M^i]$. For notational simplicity, we skip u and v in the right side of this equation. We compute the probability mass function of $\mathcal{U}$ given $\mathcal{V}$ as,

$$P_{U|V}(u_j|v_i) = \frac{\phi_j^i}{\sum_{k=1}^{M} \phi_k^i} \tag{6}$$

We can also compute the uncertainty $H(U|V = v_i)$ from the distribution of $P_{U|V}$. Please note that we have N such distributions for N classes of $\mathcal{V}$. These distributions will be used to compute the uncertainty of a sequence.

We use the predicted labels obtained from baseline classifiers to construct a sequence. As mentioned above, the instances belonging to classification task $\mathcal{U}$ can appear multiple times with instances in task $\mathcal{V}$. We develop a sequence based on the labels of $\mathcal{U}$ and use the distribution (Eqn. 6) depending on the label of the instances belonging to task $\mathcal{V}$. Let us consider that Q samples of task $\mathcal{U}$ co-occur with task $\mathcal{V}$ in an image or video. So, our sequence will be $l_1, l_2, \ldots, l_Q$, where l_p represents the predicted class given the p^{th} sample. We also know the label of $\mathcal{V}$ provided by the baseline classifier. If the label predicted by $\mathcal{V}$ classifier is v_j, then we can compute the typicality score using the distribution $P_{U|V}(u|v_j)$ for the sequence formed by $\mathcal{U}$ classifier. As we can see that the labels are provided by classifiers and the distributions are computed from the contextual relations between $\mathcal{U}$ and $\mathcal{V}$, typicality is a useful tool to inter-link between recognition and context models efficiently.

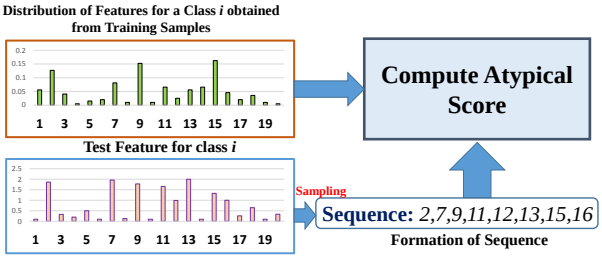


Figure 4: The figure shows how typicality can be used in feature space.

Let us look at an example. In joint scene-object classification, typicality connects all the appearing objects with scene. The object labels provided by detectors are used to construct the sequence and the label of scene or activity is used to determine the distribution to be used to compute the *atypical score* to be discussed next. Fig. 3 shows an example of how detected objects are represented as a sequence in joint scene-object scenario. Given an image, the detected objects can be represented as a sequence, where each element of the sequence is referred as 'symbol'. Q is the length of a sequence, which is the number of detected objects in an image and it varies for different images. Besides, same object labels can appear multiple times in a sequence as shown in Fig. 3.

Given an unlabeled instance (e.g. an image for scene-object or video for activity-object), we obtain the predicted labels for both tasks $\mathcal{U}$ and $\mathcal{V}$. We use the distribution $P_{U|V}$ to compute $-\log P_{U^Q|V}(\boldsymbol{u}^Q|v_i)$ for a sequence using Eqn. 1 and also compute $Q.H(U|V = v_i)$. We measure the deviation $\mathcal{D}$ from $P_{U|V}$ as,

$$\mathcal{D} = -Q.H(U|V = v_i) - \log_2 P_{U^Q}(\boldsymbol{u}^Q|v_i) \tag{7}$$

Intuitively, Eqn. 7 finds how much a sequence deviates from true distribution of co-occurrence of two tasks. We call this deviation *atypical score*. Please note that order of labels of the sequence does not affect the atypical score. A sample will be more likely to be selected for manual labeling when the score is high. In other words, we focus on the samples of a sequence that lie outside the range, represented by $|\mathcal{D}| \leq \epsilon$ (derived from Eqn. 3). Here, ϵ is a threshold by which we can determine when the properties of typicality break.

Typicality in Feature Space of Individual Class. Let us consider $\mathcal{F}^i(k)$ to be a feature vector for k^{th} sample belonging to class c_i with dimension $\mathbb{R}^{N_f \times 1}$. N_f represents number of features used in classification (e.g. the dimension of CNN feature, $N_f = 4096$). If we have N_c^i number of samples belonging to class c_i, then we can compute the mean of their feature vectors as $\hat{\mathcal{F}}^i = \frac{1}{N_c^i} \sum_{k=1}^{N_c^i} \mathcal{F}^i(k)$. Now, we can obtain the distribution by using softmax function as follows:

$$P_{\hat{F}_l|c_i}(\hat{f}_l|c_i) = \frac{\exp(\hat{\mathcal{F}}_l^i)}{\sum_{m=1}^{N_f} \exp(\hat{\mathcal{F}}_m^i)} \tag{8}$$

where $\hat{\mathcal{F}}_l^i$ represents the l^{th} element (feature value) of $\hat{\mathcal{F}}^i$. $P_{\hat{F}_l|c_i}$ denotes the distribution of average feature values of $\hat{\mathcal{F}}^i$ for class c_i with dimension N_f.

Given a test feature vector, we generate a sequence of length Q by sampling from it with replacement. This may lead to the same feature occurring multiple times in the sequence. Fig. 4 shows how a sequence is generated in feature space. Let us consider a test feature vector $F_t(\in \mathbb{R}^{N_f \times 1})$, and Q is the length of a sequence. An element of a sequence can take the value in between 1 to N_f. We assume that elements of the feature vector are independent of each other. We extract features from the final layer of CNN, where units of the layer are not internally connected between them.

As we know the sequence and the distribution, we can calculate $-\log P_{\hat{F}^P}(\hat{f}^P|c_i)$ for a sequence using Eqn. 1. Now, the deviation $\mathcal{D}_f$ can be expressed as,

$$\mathcal{D}_f = -Q.H(\hat{F}|C_i) - \log_2 P_{\hat{F}^P}(\hat{f}^P|c_i) \tag{9}$$

Here, H is calculated from $P_{\hat{F}_l|c_i}(.)$ as shown in Eqn. 8. The atypical score (or deviation) $\mathcal{D}_f$ will be used to derive an optimization function to select the most informative samples in active learning. Intuitively, we measure how much the features of a test sample deviate from its average feature values learned from training samples.

4. Active Learning Framework

We use atypical score obtained from Eqns. 9 and 7 to formulate an objective function to select the samples that need to be labeled. We first explain how the contextual relationships are used, and then combine with the feature descriptors.

Exploiting Contextual Relationships in Sample Selection. We are looking for the symbols of a sequence that lead to the sequence deviating from the typical set. In order to do that, we introduce a notation $\mathcal{D}_{q'}$ that represents the atypical score without considering a sample with index q' of the sequence. $\mathcal{D}_{q'}$ (using Eqn. 7) can be written as,

$$\begin{aligned}\mathcal{D}_{q'} &= -\log_2 P_{U^{Q-1}|V}(\boldsymbol{u}^{Q-1}|v_i) - (Q-1)H(U|V=v_i) \\ &= -\sum_{\substack{m=1 \\ m\neq q'}}^{Q} \log_2 P_{U_m|V}(u_m|v_i) - (Q-1)H(U|V=v_i)\end{aligned}$$

Here, U_m is the m^{th} symbol of a sequence. Now, we compare the $\mathcal{D}_{q'}$ with $\mathcal{D}$ (from Eqn. 7) to find the most informative samples. The difference between these two terms will be denoted as $\Delta\mathcal{D}_{q'}$ that measures how much deviation has occurred due to the symbol q'. It can be written as

$$\begin{aligned}\Delta\mathcal{D}_{q'} &= \mathcal{D} - \mathcal{D}_{q'} \\ &= -\sum_{m=1}^{Q} \log_2 P_{U_m|V}(u_m|v_i) - Q.H(U|V=v_i) \\ &+ \sum_{\substack{m=1 \\ m\neq q'}}^{Q} \log_2 P_{U_m|V}(u_m|v_i) + (Q-1)H(U|V=v_i) \\ &= -\log_2 P_{U_{q'}|V}(u_{q'}|v_i) - H(U|V=v_i)\end{aligned} \tag{10}$$

Here, $H(U|V=v_i)$ is constant for all samples. Thus, we neglect this term and focus on the first term only. We choose the sample for manual labeling by maximizing $\Delta\mathcal{D}_{q'}$, which can be represented as

$$q^* = \arg\max_{q'} \; -\log_2 P_{U_{q'}|V_i}(u_{q'}|v_i) \tag{11}$$

Thus, a sample belonging to classification task $\mathcal{U}$ will be selected if the probability of co-occurrence with the corresponding sample belonging to task $\mathcal{V}$ is very low.

Let us take the example of scene-object classification. Intuitively, even though a detector is certain about an object sample, it could be selected due to contextual irregularity. It might be possible that the detector is correct, which means that context model has not encountered the detected object much for a particular scene. But, this sample is critical to update either context model or recognition model.

Formulation of Overall Objective Function. Let us define a vector $\boldsymbol{\mathcal{T}}_f = [\boldsymbol{\mathcal{D}}_{f_1} \quad \boldsymbol{\mathcal{D}}_{f_2} \quad \ldots]^T$, which contains the atypical score of each sample using Eqn. 9. $\boldsymbol{\mathcal{D}}_{f_j}$ represents the atypical score for j^{th} sample. Similarly, for the contextual information, we consider a vector $\boldsymbol{\mathcal{T}}_j(\in \mathbb{R}^{(Q+1)\times 1})$ (e.g. Q is the number of detected objects in scene-object or activity-object classification) for the j^{th} sample in a batch of data. Please note that in Eqn. 11, there is no information of task $\mathcal{V}$ as the sequence only includes samples of $\mathcal{U}$. So, we consider $\mathcal{D}$ (as shown in Eqn. 7) in objective function as it provides global information for task $\mathcal{V}$ (e.g. scene for scene-object recognition and activity for activity-object recognition).$\boldsymbol{\mathcal{T}}_j$ can be written as follows,

$$\begin{aligned}\boldsymbol{\mathcal{T}}_j = [\mathcal{D}, &-\log_2 P_{U|V}(u_1|v_i), -\log_2 P_{U|V}(u_2|v_i), \ldots, \\ &-\log_2 P_{U|V}(u_Q|v_i)]^T\end{aligned} \tag{12}$$

$$\boldsymbol{\mathcal{T}} = [\boldsymbol{\mathcal{T}}_1 \quad \boldsymbol{\mathcal{T}}_2 \quad \ldots]^T \tag{13}$$

It may be noted that the elements of the vector $\boldsymbol{\mathcal{T}}_j$ should also have an index j as the vector is different for different samples (images or videos), but for the sake of simplicity, we have dropped the index j from its elements.

We also involve the uncertainty of the current baseline classifier on the unlabeled samples to choose the informative samples. We define a vector of the entropy of the samples as, $\boldsymbol{h} = [h_1 \quad h_2 \quad \ldots]^T$, where $h_j = \mathbb{E}[-\log_2 p_k]$, p_k is the p.m.f. of prediction by the current baseline classifier on the k^{th} unlabeled instance. 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. We can write the optimization function in vector form as follows,

$$\begin{aligned}\boldsymbol{y}^* &= \arg\max_{\boldsymbol{y}} \; \boldsymbol{y}^T(\boldsymbol{h} + \lambda_1\boldsymbol{\mathcal{T}}_f + \lambda_2\boldsymbol{\mathcal{T}} - \beta\mathbf{1}) \\ s.t. &\quad \boldsymbol{y} \in \{0,1\}^N, \quad (\mathbf{1}-\boldsymbol{y})^T\boldsymbol{h} \leq \xi\end{aligned} \tag{14}$$

Here, λ_1, λ_2 and β are weighting factors. The additional term $\boldsymbol{y}^T\mathbf{1}$ weighted by β in the objective function tries to

minimize the total number of selected samples. Let us denote $\boldsymbol{f} = -(\boldsymbol{h} + \lambda_1 \boldsymbol{\mathcal{T}}_f + \lambda_2 \boldsymbol{\mathcal{T}} - \beta \mathbf{1})$. Maximization of the objective function in Eqn. 14 is the same as minimization of $\boldsymbol{y}^T \boldsymbol{f}$, which is now a convex optimization problem. It is a binary linear integer programming and can be solved by CPLEX [5]. After obtaining a set of samples $\boldsymbol{y}^*$ from Eqn. 14, we can ask a human to label these samples.

Classifier Update. In this paper, 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 [6]. At current batch, we update the parameters with newly labeled data samples.

PMF Update in Typicality. The co-occurrence statistics $\Phi(u|v_i)$ are updated based on the newly acquired labels of tasks $\mathcal{U}$ and $\mathcal{V}$. The updated statistics can be written as, $\Phi'(u|v_i) \leftarrow \Phi(u|v_i) + \tilde{\Phi}(u|v_i)$, where $\tilde{\Phi}(.)$ represents the statistics with the newly labeled samples and Φ' is the updated statistics. Similarly, we also update $\hat{\mathcal{F}}_l^i$ used in Eqn. 8 with newly labeled data.

5. Experiments

We perform both image and video classification tasks to evaluate our proposed sample selection framework. We also demonstrate our results on joint classification tasks such as scene-object and activity-objects classification, where contextual relationships between samples are exploited.

Experimental Setup. We consider an online setting, where samples (e.g. images or videos) are continuously coming in batches. Batches are generated from training set and results are evaluated on testing set. We use the samples from the first batch to build the initial model as well as context model. We also incorporate incremental learning to update the model as new classes can come in new batches. We always use current batch of data to update the previous recognition model.

Evaluation Criterion. We obtain the recognition accuracy by using SVM classifier for scene and activity classification. For object detection, we compute the average precision by comparing with the ground truth. We consider intersection over ratio (IoU) between the detected box and the ground-truth bounding box to localize an object. IoU ratio, greater than equal to 50%, is considered as correct detection.

Datasets. In joint scene-object classification, we use MSRC [33] and MIT-67 Indoor [38] datasets to evaluate the propose framework. These datasets are appropriate as they provide a rich source of contextual information between scene and objects. For MSRC [33] dataset, we use all the classes to compute recognition accuracy with the ground truth provided by [47]. For MIT-67 indoor [38] dataset, we use 67 scene categories and 50 object categories. We use CAD-120[24] to evaluate results on joint activity-objects classification. We also demonstrate our results on Scene-15 [26] and VOC2010 [11] datasets for scene classification and object recognition results. Scene-15 and VOC2010 datasets only provide the ground-truth for one classification task, thus we do not consider any contextual relations ($\lambda_2 = 0$, in Eqn. 14) in the experiments.

Baseline Methods: In the experiment, we use following baseline methods.

◇ **Typicality^1:** Proposed framework, the recognition accuracy is obtained from the baseline classifier.

◇ **Typicality^2:** Proposed framework where the accuracy is obtained from the marginal(posterior) probability of a graph by exploiting contextual relationship. The prior probabilities of a graph are provided by the baseline classifiers.

◇ **SOAL:** Scene-object active learning (SOAL) [2].

◇ **Bv2B:** Best vs Second Best active learning strategy [28].

◇ **IL:** Incremental learning approach presented in [17].

◇ **Full-set^2:** Entire training set with graph is considered.

◇ **Full-set^1:** Entire training is used to obtain the accuracy from baseline classifiers.

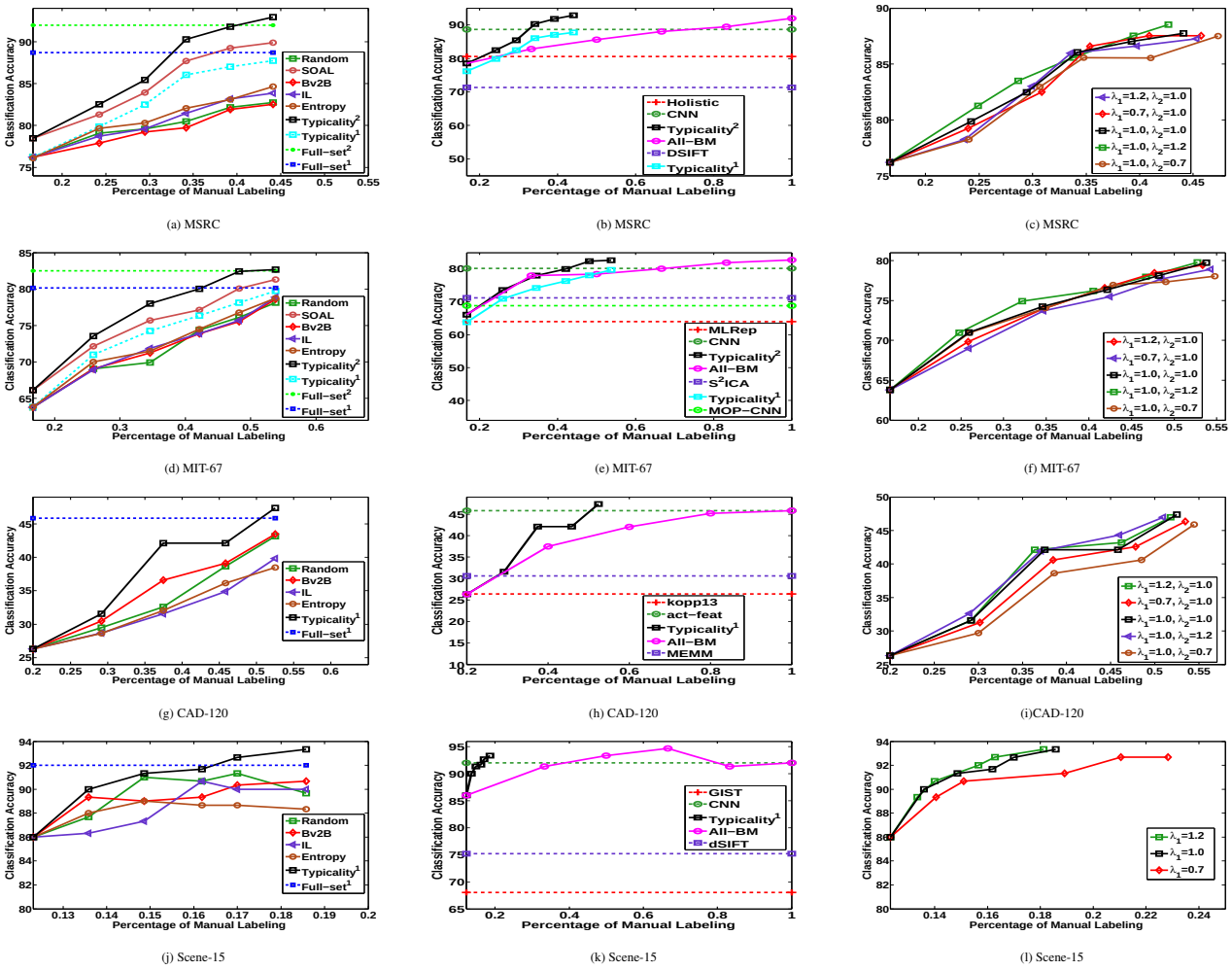
◇ **BM-All:** All the samples in current batch are considered.

Feature Extraction. In scene classification and object recognition, we consider the CNN features from image and region proposals [15] respectively. For CNN feature, N_f discussed in Sec. 3 would be 4096. For activity recognition, we consider the features provided in [24] with dimension, $N_f = 630$. We refer to these features as *'act-feat'*.

Experimental Analysis: We perform the following set of experiments - 1. Comparison with other active learning methods, 2. Comparison against other recognition methods, and 3. Sensitivity analysis of the parameters.

Comparison With Other Active Learning Methods. We compare our active learning (AL) framework with other state-of-the-art methods and baseline approaches as shown in Figs. 5(a,d,g,j) and 6(a,d,g,j). The straight line presented in the figures implies the recognition accuracy on whole training set. Some of the existing AL approaches are SOAL [2], Bv2B [28], random sample selection, Entropy [9] and IL [17]. We observe the recognition accuracy as a function of number of samples chosen by proposed method. Then, we fix the number of samples for each batch, and obtain the accuracy for other AL methods. Here, different methods choose different set of samples, from which recognition models are trained. Features and baseline classifiers are kept same for all the methods for fair comparison. From Figs. 5(a,d,g,j) and 6(a,d,g,j), we can see that the proposed framework *outperforms other methods by large margin in selecting the most informative samples* in all the classification tasks- scene, object and activity classification.

Comparison Against Other Classification Methods. We compare our framework against other state-of-the-art recognition methods. We implement some of the methods-CNN [50], GIST, DSIFT [31], R-CNN [15], DPM [13] for scene and object classification. In scene classifica-



(a) MSRC (b) MSRC (c) MSRC

(d) MIT-67 (e) MIT-67 (f) MIT-67

(g) CAD-120 (h) CAD-120 (i)CAD-120

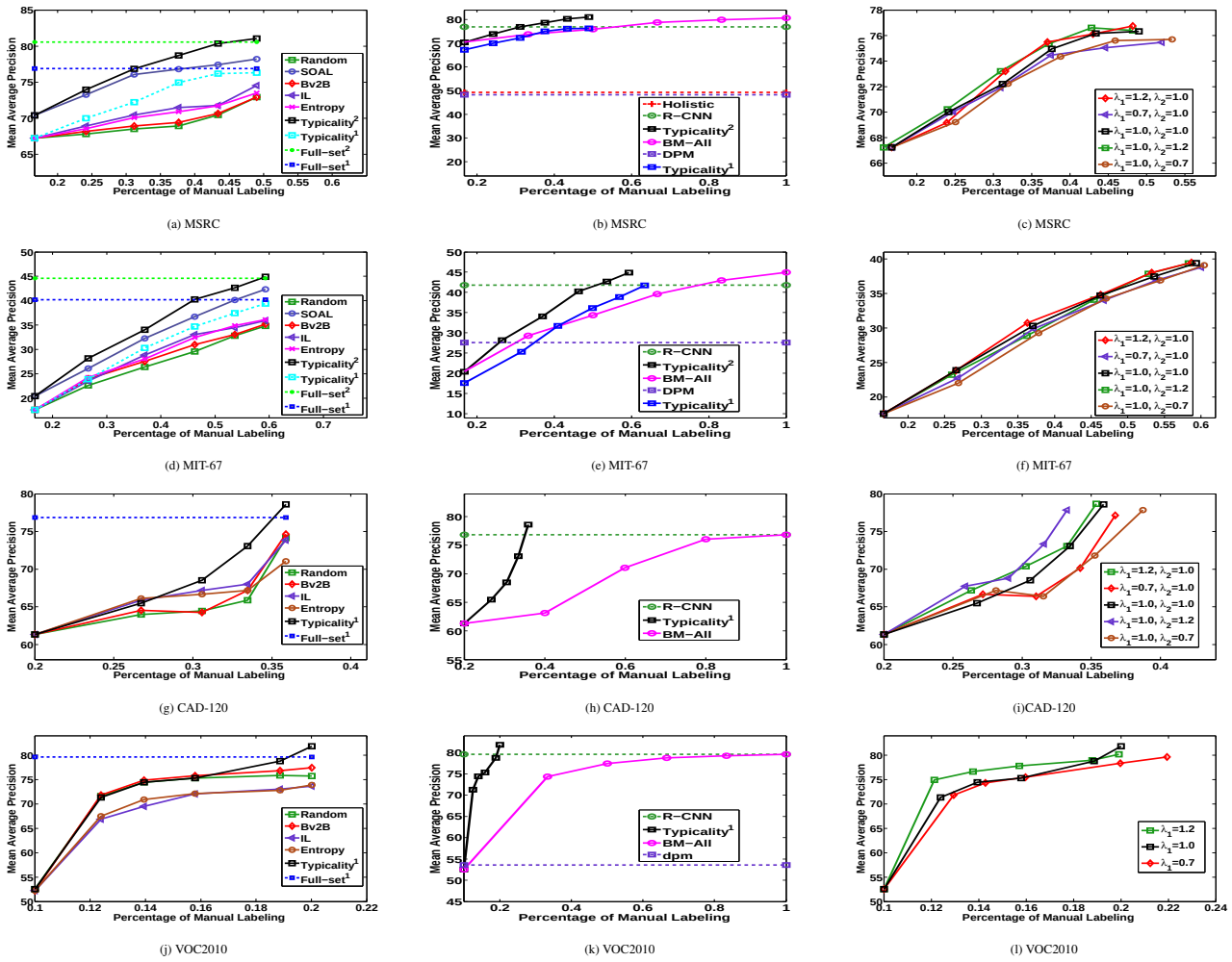
(j) Scene-15 (k) Scene-15 (l) Scene-15

Figure 5: The figure presents scene and activity classification performance for four datasets- MSRC [33], MIT-67 [38], CAD120-activity [24] and Scene-15 [26] (top to down). Plots (a,d,g,j) present the comparison against other state-of-the-art active learning methods. Plots (b,e,h,k) demonstrate comparison with other recognition methods. Plots (c,f,i,l) demonstrate the sensitivity analysis of our framework.

tion, we also compare against Holistic [47], MLRep [8], S^2ICA [19] and MOP-CNN [16] methods. Similarly, we also consider Holistic [47] approach for object detection performance. For activity recognition, we compare against MEMM [42], Kopp13[24] methods. The recognition performance is shown in Fig. 5(b,e,h,k) for scene and activity classification, and object detection performance is shown in Fig. 6(b,e,h,k). We also include the plot for BM-ALL methods to illustrate the impact of the proposed method in selecting the most informative samples. BM-ALL represents all the samples in a current batch, thus for n batches we have n accuracy values. It is obvious that selection of the informative samples plays an important role in adapting a recognition model. In Fig. 6(h,k), with small number of samples, our method demonstrates similar performance compared to BM-ALL method. Figs. 5(b,e,h,k) and 6(b,e,h,k) demonstrate that the proposed framework performs better with fewer informative samples when compared to the other recognition models.

Sensitivity Analysis of the Parameters. In our active learning framework, even though we use parameters λ_1, λ_2, β and ξ in Eqn. 14, we show our results with varying λ_1 and λ_2 as these parameters are associated with typicality. We see the effect of $\mathcal{T}_f$ and $\mathcal{T}$ as presented in Eqn. 14 in selecting the most informative samples. Towards this goal, we choose the values of λ_1 and λ_2 as $0.7, 1.0, 1.2$. $\lambda_2 = 0$ for Scene-15 and VOC2010 datasets, which means that no contextual relation is used. Figs. 5(c,f,i,l) and 6(c,f,i,l) illustrate the variation of performance due to change in parameters. From figures, we can see that the performance is improved when we have more weight to emphasize typicality.

Computational Complexity. We analyze the complexity in terms of computational time on MSRC [33] and MIT-67 [38] datasets. We compute the time to query the samples, and time to train scene and object models for a dataset. As we can see that total time to train scene and object



(a) MSRC (b) MSRC (c) MSRC

(d) MIT-67 (e) MIT-67 (f) MIT-67

(g) CAD-120 (h) CAD-120 (i)CAD-120

(j) VOC2010 (k) VOC2010 (l) VOC2010

Figure 6: In this figure, we show the object detection performances on MSRC [33], MIT-67 [38],CAD120-activity [24] and Scene-15 [26] (top to down). Plots (a, d,g,j) present the comparison of other state-of-the-art active learning methods. Plots (b, e, h,k) demonstrate comparison with different recognition techniques. Plots (c,f,i,l) present the sensitivity analysis of the proposed framework.

Dataset	QT	Train SM (s)		Train OM (s)	
	(s)	with SS	all	with SS	all
MSRC	19.72	42.17	47.58	359.95	657.02
MIT-67	63.07	113.52	384.91	1187.9	1775.1

Table 1: Timing analysis on MSRC [33], MIT-67 [38] datasets. Here, **SM**-scene model, **OM**-object model, **QT**-query time, **SS**-selected samples

models with all the samples is $704.58s(47.58 + 657)$ for MSRC [33] and $2160.01s(384.91 + 1755.1)$ for MIT-67 [38]. On the other hand, total time for querying and training with samples selected by our approach is $421.84s(19.72 + 42.17 + 359.95)$ and $1364.59s(63.09 + 113.52 + 1187.9)$ for MSRC [33], and MIT-67 [38] respectively. We can conclude that the proposed AL method will help saving significant amount of computational time, especially in big dataset.

6. Conclusions

In this paper, we propose a novel subset selection framework to adaptively learn the recognition models. We introduce the typicality concept which can be used as an important tool to learn informative samples from a huge pool of unlabeled samples. We efficiently link between recognition and context model by exploiting typicality. We can also apply typicality in feature space to learn a good recognition model. Our approach significantly reduces the load on human effort in labeling samples. We also show that with only a small subset of the full training set we achieve better or similar performance compared with using full training set.

Acknowledgment. This work was partially funded by NSF grant IIS-1316934 from the National Robotics Initiative.

References

[1] M. Alberti, J. Folkesson, and P. Jensfelt. Relational approaches for joint object classification and scene similarity measurement in indoor environments. In *AAAI 2014 Spring Symposia: Qualitative Representations for Robots*, 2014. 2

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

[3] W. Choi, K. Shahid, and S. Savarese. Learning context for collective activity recognition. In *CVPR*, pages 3273–3280, 2011. 2

[4] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012. 3

[5] I. I. CPLEX. V12. 1: Users manual for cplex. *International Business Machines Corporation*, 46(53):157, 2009. 6

[6] 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. 6

[7] J. Deng, O. Russakovsky, J. Krause, M. S. Bernstein, A. Berg, and L. Fei-Fei. Scalable multi-label annotation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3099–3102. ACM, 2014. 3

[8] C. Doersch, A. Gupta, and A. A. Efros. Mid-level visual element discovery as discriminative mode seeking. In *NIPS*, 2013. 7

[9] G. Druck, B. Settles, and A. McCallum. Active learning by labeling features. In *EMNLP*, 2009. 6

[10] E. Elhamifar, G. Sapiro, A. Yang, and S. Sasrty. A convex optimization framework for active learning. In *ICCV*, 2013. 3

[11] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010. 6

[12] A. Fathi, M. F. Balcan, X. Ren, and J. M. Rehg. Combining self training and active learning for video segmentation. In *BMVC*, volume 29, pages 78–1, 2011. 3

[13] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627–1645, 2010. 6

[14] R. Girshick. Fast r-cnn. In *ICCV*, 2015. 2

[15] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 2, 6

[16] Y. Gong, L. Wang, R. Guo, and S. Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *ECCV*. 2014. 7

[17] M. Hasan and A. Roy-Chowdhury. Incremental activity modeling and recognition in streaming videos. In *CVPR*, 2014. 6

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

[19] M. Hayat, S. H. Khan, M. Bennamoun, and S. An. A spatial layout and scale invariant feature representation for indoor scene classification. *arXiv preprint arXiv:1506.05532*, 2015. 7

[20] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV 2014*, pages 346–361. 2014. 2

[21] X. Hu, J. Tang, H. Gao, and H. Liu. Actnet: Active learning for networked texts in microblogging. In *SDM*, pages 306–314. SIAM, 2013. 3

[22] C. Kading, A. Freytag, E. Rodner, P. Bodesheim, and J. Denzler. Active learning and discovery of object categories in the presence of unnameable instances. In *CVPR*, 2015. 3

[23] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell. Active learning with gaussian processes for object categorization. In *ICCV*, 2007. 3

[24] H. S. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 2013. 6, 7, 8

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

[26] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 6, 7, 8

[27] J. Li, J. M. Bioucas-Dias, and A. Plaza. Spectral–spatial classification of hyperspectral data using loopy belief propagation and active learning. *Geoscience and Remote Sensing, IEEE Transactions on*, 51(2):844–856, 2013. 1, 3

[28] X. Li, R. Guo, and J. Cheng. Incorporating incremental and active learning for scene classification. In *ICMLA*, 2012. 3, 6

[29] X. Li and Y. Guo. Adaptive active learning for image classification. In *CVPR*, 2013. 1, 3

[30] X. Li and Y. Guo. Multi-level adaptive active learning for scene classification. In *ECCV*. 2014. 3

[31] C. Liu, J. Yuen, and A. Torralba. Dense scene alignment using sift flow for object recognition. In *CVPR*, 2009. 6

[32] O. Mac Aodha, N. Campbell, J. Kautz, and G. Brostow. Hierarchical subquery evaluation for active learning on a graph. In *CVPR*, 2014. 1, 3

[33] T. Malisiewicz and A. A. Efros. Improving spatial support for objects via multiple segmentations. In *BMVC*, 2007. 6, 7, 8

[34] J. T. Maxfield, W. D. Stalder, and G. J. Zelinsky. Effects of target typicality on categorical search. *Journal of vision*, 14(12):1–1, 2014. 2

[35] T. Nimmagadda and A. Anandkumar. Multi-object classification and unsupervised scene understanding using deep learning features and latent tree probabilistic models. *arXiv preprint arXiv:1505.00308*, 2015. 2

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

[37] R. Poppe. A survey on vision-based human action recognition. *Image and vision computing*, 28(6):976–990, 2010. 2

[38] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, 2009. 6, 7, 8

[39] B. Saleh, A. Elgammal, and J. Feldman. The role of typicality in object classification: Improving the generalization capacity of convolutional neural networks. *arXiv preprint arXiv:1602.02865*, 2016. 2

[40] B. Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66), 2010. 1, 3

[41] B. Settles. Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114, 2012. 1, 3

[42] J. Sung, C. Ponce, B. Selman, and A. Saxena. Unstructured human activity detection from rgbd images. In *ICRA*, 2012. 7

[43] D. P. Tian. A review on image feature extraction and representation techniques. *International Journal of Multimedia and Ubiquitous Engineering*, 8(4):385–396, 2013. 2

[44] S. Vijayanarasimhan and K. Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. *IJCV*, 108(1-2):97–114, 2014. 3

[45] J. Vogel and B. Schiele. A semantic typicality measure for natural scene categorization. In *Joint Pattern Recognition Symposium*, pages 195–203. Springer, 2004. 2

[46] C. Vondrick and D. Ramanan. Video annotation and tracking with active learning. In *NIPS*, 2011. 3

[47] J. Yao, S. Fidler, and R. Urtasun. Describing the scene as a whole: Joint object detection, scene classification and semantic segmentation. In *CVPR*, 2012. 2, 6, 7

[48] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory*, 51(7):2282–2312, 2005. 1

[49] L. Zhang, X. Zhen, and L. Shao. Learning object-to-class kernels for scene classification. *TIP*, 23(8):3241–3253, 2014. 2

[50] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *NIPS*, pages 487–495, 2014. 2, 6