Uncertainty Sampling for Action Recognition via Maximizing Expected Average Precision

Hanmo Wang[1,2], **Xiaojun Chang**[3], **Lei Shi**[1,2], **Yi Yang**[4], **Yi-Dong Shen**[1*]
[1] State Key Lab. of Computer Science, Institute of Software, Chinese Academy of Sciences, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] School of Computer Science, Carnegie Mellon University, Pittsburgh, USA
[4] Centre for Artificial Intelligence, University of Technology Sydney, Sydney, Australia
{wanghm, shilei, ydshen}@ios.ac.cn, cxj273@gmail.com, yi.yang@uts.edu.au

Abstract

Recognizing human actions in video clips has been an important topic in computer vision. Sufficient labeled data is one of the prerequisites for the good performance of action recognition algorithms. However, while abundant videos can be collected from the Internet, categorizing each video clip is time-consuming. Active learning is one way to alleviate the labeling labor by allowing the classifier to choose the most informative unlabeled instances for manual annotation. Among various active learning algorithms, uncertainty sampling is arguably the most widely-used strategy. Conventional uncertainty sampling strategies such as entropy-based methods are usually tested under accuracy. However, in action recognition Average Precision (AP) is an acknowledged evaluation metric, which is somehow ignored in the active learning community. It is defined as the area under the precision-recall curve. In this paper, we propose a novel uncertainty sampling algorithm for action recognition using expected AP. We conduct experiments on three real-world action recognition datasets and show that our algorithm outperforms other uncertainty-based active learning algorithms.

1 Introduction

Recognizing human actions in video clips has been an important topic itself, and also as a component in more complex tasks such as event detection. Most action recognition datasets are collected from various sources, including Youtube and Google videos, Hollywood movies, and so on. With sufficient labeled data, action recognition algorithms are able to achieve good performance. However, while collecting the videos takes effort, labeling them is a more complex task in terms of time and money. Active learning is one way to alleviate the pain of annotation by allowing the classifier to choose the most informative data samples for labeling. It has many applications in computer vision such as segmentation [Dutt Jain and Grauman, 2016], pose estimation [Liu and Ferrari, 2017], and event detection [Yang *et al.*, 2015].

When it comes to applying active learning in classification tasks, uncertainty sampling is arguably the "first choice" owing to its simplicity and effectiveness. However, there is no guarantee that the traditional task-agnostic uncertainty measure matches the evaluation metric of the task at hand. In addition, recent advance [Ramirez-Loaiza *et al.*, 2017] in active learning suggests that conventional uncertainty sampling algorithms may not work when the underlying evaluation metric is *not* accuracy. Therefore, we argue that the uncertainties of data samples change along with the underlying evaluation metric. In action recognition, Average Precision (AP) is a widely-used evaluation metric that calculates the area under the precision-recall curve, which is somehow overlooked in the active learning community. This paper aims at resolving the issues with the conventional uncertainty sampling algorithms by understanding their intrinsic properties, and developing new active learning algorithms for action recognition.

We propose to evaluate the uncertainty of one video by the expected AP on all other unlabeled videos. Similar to Expected Error Reduction [Roy and McCallum, 2001], videos with maximal expected AP are selected for labeling. Unlike accuracy which can be re-interpreted as the average of individual scores over all samples, AP is more related to complex rank-based metrics (such as AUC). It cannot be viewed as the average of instance-level scores and thus cannot be directly calculated using previous techniques such as [Roy and McCallum, 2001]. In order to obtain expected AP, we face two inevitable obstacles. The first one is "how to calculate AP when the *true labels* of video clips are unknown," while the

[*] Corresponding author

second one is "under what *distribution* is the expected AP-based calculated."

Although the *true label* of each unlabeled video clip is unknown, the probability that the video belongs to the positive (negative) class is provided by the underlying one-vs-rest classifiers. We treat each unknown *true label* as a independent biased coin that flips head (one) with probability to the positive class, and tail (zero) with probability to the negative class. In other words, the true label associated with each unlabeled video clip is defined as an independent Bernoulli random variable, and the probability of success is defined as the probability that the video belongs to the positive class. After that, we explicitly calculate the expected AP on the *joint distribution* of all true labels. Since most action recognition datasets have more than two classes, the expected AP is calculated in a one-vs-rest manner. To the best of our knowledge, this is the first work that optimizes AP in active learning for classification. It is also suitable for action recognition tasks, since AP is widely used in such applications.

Our contributions are as follows. First, we introduce a novel technique that treats the true labels of unlabeled videos as independent Bernoulli random variables, thus making it possible to directly calculate the expected evaluation metric. Second, this technique successfully reveals the intrinsic properties of existing uncertainty sampling algorithms. Third, we calculate the expected AP using dynamic programming in polynomial time, and empirically evaluate our algorithm on three real-world action recognition datasets, and show that our uncertainty sampling approach outperforms other uncertain-based methods with (mean) AP being the evaluation metric.

2 Related Work

There is an extensive body of literature about action recognition; here we only name a few related to the feature extractor used in this paper. Trajectory-based methods such as Dense Trajectories (DT) [Wang *et al.*, 2011] and Improved Dense Trajectories (IDT) [Wang and Schmid, 2013] are among the most effective feature extractors in action recognition. There are many works in literature that tries to improve on IDT, such as multi-skip feature stacking [Lan *et al.*, 2015] and Fisher Vector encoding [Perronnin *et al.*, 2010]. Since AP is an important evaluation metric in many tasks including action recognition, researchers propose different models that optimize AP during training. For example, [Triantafillou *et al.*, 2017] optimizes AP for information retrieval models. [Yue *et al.*, 2007] proposes a variant of the SVMs that optimizes AP. [Behl *et al.*, 2014] further develops this method in weakly supervised learning by minimizing an upper-bound for AP .

Pool-based active learning is an intensively-studied problem in the machine learning community (See [Settles, 2010] for an overview). In the past two decades, many active learning algorithms are developed, such as selecting uncertain samples [Lewis and Gale, 1994], selecting samples about which a committee of classifier disagree [Seung *et al.*, 1992], querying samplings that minimizes the expected error of the model [Roy and McCallum, 2001], selecting representative samples [Yu *et al.*, 2006], and many more. One of the most widely-used groups of active learning algorithms is uncertainty sampling [Lewis and Gale, 1994; Sharma and Bilgic, 2017], which selects the instances about which the classifier is most uncertain for labeling. Uncertainty sampling is widely used in computer vision tasks such as medical image classification [Zhou *et al.*, 2017], human pose estimation [Liu and Ferrari, 2017], facial expression recognition [Chakraborty *et al.*, 2015], action recognition/event detection [Yang *et al.*, 2015], and so on. In most active learning literature, accuracy is chosen as the evaluation metric, but researchers also develop active learning algorithms for other metrics. For example, [Culver *et al.*, 2006] proposes a method that maximizes AUC of the hypothesis, using a semi-supervised ranking approach. [Long *et al.*, 2010] maximizes discounted cumulative gain (DCG) to select the most informative instances. To the best of our knowledge, there is currently no uncertainty sampling algorithm that directly optimizes AP in literature.

3 The Proposed Method

In this section, we first review two most widely-used uncertainty measures, namely maximum conditional and entropy, and then we propose to treat unknown true labels as Bernoulli random variables and show that maximum conditional implicitly optimizes expected accuracy. After that, we propose to optimize expected AP and obtain our Uncertainty Sampling via maximizing expected Average Precision (*USAP*).

3.1 Uncertainty Sampling

Typical uncertainty sampling strategies select top-k instances from the unlabeled pool U based on their individual uncertainty $\phi(\cdot)$, formally we have

$$S^* = \underset{S\subset U, |S|=k}{\arg\max} \sum_{x\in S} \phi(x)$$

where S^* is the set of the most uncertain samples. Two common uncertainty measures are maximum conditional and Shanon entropy. For binary classification, we have

$$\phi_{condi}(x) = -\max(P_1, 1-P_1) \tag{1}$$

$$\phi_{entro}(x) = -P_1 log P_1 - (1-P_1) log(1-P_1) \tag{2}$$

where P_1 indicates the probability sample x belong to the positive class. Note that Eq. (1) and Eq. (2) selects the exact same samples in binary classification. There are also multi-class versions of the two algorithms

$$\phi_{condi}(x) = \max_y -P(y|x,w) \tag{3}$$

$$\phi_{entro}(x) = \sum_y -P(y|x,w) log P(y|x,w) \tag{4}$$

where $P(y|x,w)$ indicates the probability that instance x belongs to class y given the classifier with weight w.

3.2 Uncertainty Sampling under Accuracy

For binary classification, accuracy is defined as the proportion of samples that are correctly classified over all tested samples, i.e., $accuracy = (tp + tn)/(tp + fp + tn + fn)$. For n unlabeled samples $\{x_i\}_{i=1}^n$, let the labels be selected from $\{0, 1\}$ and let p_i be the posterior probability that x_i belongs to the positive class, i.e., $p_i = P(y = 1|x_i, w)$. Let $\{y_i\}_{i=1}^n$ be a set of independent Bernoulli random variables such that y_i takes one with probability p_i and zero otherwise. In other words, we have

$$y_i \sim Ber(p_i) \tag{5}$$

Intuitively, for each unlabeled instance x_i, we flip a biased coin and let the outcome y_i be the true labels of x_i. Given the true labels, the accuracy can be easily calculated as follows

$$ACC = \frac{1}{n}\sum_{i=1}^{n} I(y_i = \hat{y}_i) \tag{6}$$

where $I(\cdot)$ is the indicator function and $\hat{y}_i$ is the predicted label of x_i. For binary classification, we usually assume $\hat{y}_i$ takes one if $p_i > 1 - p_i$ and zero otherwise, i.e.,

$$\hat{y}_i = I(p_i > 0.5) \tag{7}$$

By flipping n coins once, we obtain one sample of the accuracy. If the coins are flipped multiple times, the accuracies obtained each time will center around its mean. To avoid actually generating y_i for each x_i, we can calculate the expectation of accuracy over the joint distribution of all y_i, i.e.,

$$\widehat{ACC} = E_{\substack{y_i \sim Ber(p_i) \\ i=1,\ldots,n}} ACC \tag{8}$$

After substituting Eq. (6) and Eq. (7) into Eq. (8), we have

$$\begin{aligned}\widehat{ACC} &= E_{\substack{y_i \sim Ber(p_i) \\ i=1,\ldots,n}} \frac{1}{n}\sum_{i=1}^{n} I(y_i = \hat{y}_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} E_{\substack{y_i \sim Ber(p_i) \\ i=1,\ldots,n}} I(y_i = I(p_i > 0.5)) \\ &= \frac{1}{n}\sum_{i=1}^{n} p_i \times I(p_i > 0.5) + (1 - p_i) \times I(p_i \le 0.5) \\ &= \frac{1}{n}\sum_{i=1}^{n} \max(p_i, 1 - p_i)\end{aligned} \tag{9}$$

Recall that uncertainty of one instance is defined as the expected evaluation metric on all other instances, i.e., for instance x_j, we have its uncertainty $\phi(x_j)$ under accuracy as

$$\begin{aligned}\phi(x_j) &= \frac{1}{n-1}\sum_{i=1, i\neq j}^{n} \max(p_j, 1 - p_j) \\ &= \frac{1}{n-1}(\sum_{i=1}^{n} \max(p_j, 1 - p_j) - \max(p_j, 1 - p_j))\end{aligned} \tag{10}$$

After dropping constants, we have the uncertainty measure that matches the maximum conditional in Eq. (1) as follows

$$\phi_{ACC}(x_j) = -\max(p_j, 1 - p_j) = \phi_{condi}(x_j) \tag{11}$$

By introducing Bernoulli random variables as *true* labels, we prove that the maximum conditional strategy implicitly maximizes the expected *accuracy*, which is usually the underlying evaluation metric. However, recent study [Ramirez-Loaiza *et al.*, 2017] shows that when the underlying evaluation metric is not accuracy, the maximum conditional strategy may fail. Since AP is widely used in action recognition as a evaluation metric, it is desirable to select instances that directly optimizes AP. In the following, we develop an uncertainty sampling algorithm under AP using the same technique.

3.3 Uncertainty Sampling under AP

AP is a more complex evaluation metric than accuracy. It can be described as the area under the precision-recall curve. It is originally used as an evaluation metric for information retrieval tasks, and is now widely-used in classification applications such as action recognition.

For video samples $\{x_i\}_{i=1}^n$ with true labels $\{y_i\}_{i=1}^n \in \{0, 1\}^n$, let $\{p_i\}_{i=1}^n$ be the probability output of the classifier satisfying $p_i = P(y = 1|x_i, w)$. Without loss of generality, we assume that $\{p_i\}_{i=1}^n$ are sorted in descending order, i.e, $p_1 \ge p_2 \ge \ldots \ge p_n$. For convenience, we also define the positive label set $U^+ = \{i|y_i = 1\}$ and let $|U^+|$ be the cardinality of U^+. For each $i \in U^+$, by carefully setting the classification threshold so that only the first i instances are predicted positive, we have recall that equals $i/|U^+|$. The corresponding precision equals $\sum_{j=1}^{i} y_j/i$. In other words, each $i \in U^+$ corresponds to a single precision-recall point on the precision-recall curve. Therefore the area under the curve can be approximated by

$$AP = \frac{1}{|U^+|}\sum_{i\in U^+}\sum_{j=1}^{i} y_j/i \tag{12}$$

Using the fact that $|U^+| = \sum_{i=1}^{n} y_i$, and let $\{y_i\}_{i=1}^n$ be the Bernoulli random variables satisfying $y_i \sim Ber(p_i)$, the expected AP can be formulated as

$$\widehat{AP} = E_{\substack{y_i \sim Ber(p_i) \\ i=1,\ldots,n}} \frac{\sum_{i=1}^{n} y_i \sum_{j=1}^{i} y_j/i}{\sum_{j=1}^{n} y_j} \tag{13}$$

At first glance, the above expectation is difficult to calculate because the denominator $\sum_{j=1}^{n} y_j$ is not a constant. Here we expand the expectation by its definition as

$$\widehat{AP} = \sum_{\substack{y_i\in\{0,1\} \\ i=1,\ldots,n}} \prod_{l=1}^{n} p_l^{y_l}(1 - p_l)^{1-y_l} \frac{\sum_{i=1}^{n} y_i \sum_{j=1}^{i} y_j/i}{\sum_{j=1}^{n} y_j}$$

We introduce a auxiliary variable $t = \sum_{j=1}^{n} y_j$, and the expected AP becomes[1]

$$\widehat{AP} = \sum_{t=1}^{n} \frac{f(n, t)}{t} \tag{14}$$

[1] we set AP=0 when $\sum y = 0$

where

$$f(n,t) = \sum_{\substack{y_i \in \{0,1\} \\ i=1,\ldots,n}} I(t = \sum_{j=1}^{n} y_j) h(n) \tag{15}$$

and

$$h(n) = \prod_{l=1}^{n} p_l^{y_l}(1-p_l)^{1-y_l} \sum_{i=1}^{n} y_i \sum_{j=1}^{i} y_j / i$$

Therefore, instead of directly calculating expected AP, we can calculate function $h(\cdot)$ and $f(\cdot,\cdot)$ instead. For $n >= 1$ and $t >= 1$, we can calculate $h(\cdot)$ (and further $f(\cdot,\cdot)$) using dynamic programming by splitting Eq. (15) into two terms which separately indicate the case that $y_n = 0$ and $y_n = 1$. Formally, we have

$$h(n) = \underbrace{p_n(h(n-1) + \prod_{l=1}^{n-1} p_l^{y_l}(1-p_l)^{1-y_l} \sum_{j=1}^{n} y_j / n)}_{y_n=1} + \underbrace{(1-p_n)h(n-1)}_{y_n=0}$$

and thus $f(n,t)$ can be reformulated as

$$f(n,t) = \underbrace{p_n f(n-1,t-1) + \frac{p_n t}{n} g(n-1,t-1)}_{y_n=1} + \underbrace{(1-p_n)f(n-1,t)}_{y_n=0} \tag{16}$$

where

$$g(n,t) = I(\sum_{i=1}^{n} y_i = t) \sum_{\substack{y_i \in \{0,1\} \\ i=1,\ldots,n}} \prod_{l=1}^{n} p_l^{y_l}(1-p_l)^{1-y_l}$$

Similar to Eq. (16), we obtain formula for function $g(\cdot,\cdot)$ by splitting on y_n. Formally, we have

$$g(n,t) = p_n g(n-1,t-1) + (1-p_n) g(n-1,t)$$

Considering the corner cases, $g(n,t)$ $(t >= 0, n >= 0)$ has the following formulation:

$$g(n,t) = \begin{cases} 0 & t > n, \text{ or } t = 0, n > 0 \\ 1 & t = 0, n = 0 \\ \begin{array}{l} p_n g(n-1,t-1) \\ +(1-p_n) g(n-1,t) \end{array} & \text{otherwise} \end{cases} \tag{17}$$

Similarly, $f(n,t)$ $(t >= 0, n >= 0)$ can be reformulated as

$$f(n,t) = \begin{cases} 0 & t > n, \text{ or } t = 0, n > 0 \\ 1 & t = 0, n = 0 \\ \begin{array}{l} p_n f(n-1,t-1) \\ +\frac{p_n t}{n} g(n-1,t-1) \\ +(1-p_n) f(n-1,t) \end{array} & \text{otherwise} \end{cases} \tag{18}$$

Most of the existing action recognition datasets have more than two classes, therefore we use one-vs-rest classifiers and sum the expected AP over all classes. Algorithm 1 demonstrates our Uncertainty Sampling via maximizing Average Precision (USAP) method. When the number of unlabeled samples becomes large, it would be time-consuming to calculate the exact expected AP. To alleviate this issue, we use a fast screening rule to obtain fast approximations for the expected AP, and use it to filter out uninformative samples. Algorithm 1 runs in $\mathcal{O}(cn^3)$ for c-class classification with n unlabeled samples.

Algorithm 1 USAP

```
Input: number of selected videos $k$, number of classes $c$,
    number of unlabeled videos $n$, probability estimate $p \in [0,1]^{n \times c}$, unlabeled video set $U$
Output: selected video set $S$
$\phi \leftarrow \mathbf{0}$
for i=1 to c do
    for j=1 to n do
        $p' \leftarrow p_{*i} \backslash \{p_{ji}\}$% the $i$-th column of $p$ without $p_{ji}$
        Sort $p'$ in descending order
        Calculate $g(\cdot,\cdot)$ using Eq. (17) and $p'$
        Calculate $f(\cdot,\cdot)$ using Eq. (18) and $p'$
        Calculate $\widehat{AP}$ using Eq. (14)
        $\phi_j \leftarrow \phi_j + \widehat{AP}$
    end for
end for
Select $S \subset U$ corresponding to the $k$ largest $\phi$ value
Return: $S$
```

3.4 Fast Screening Rule

As the previous section shows, calculating the expected AP can be slow in practice. Therefore, we use the $precision@n$ for n video samples as an approximation to filter out uninformative samples.

Precision is the fraction of positive instances among the positively predicted instances, i.e. $PREC = tp/(tp + fn)$. Given probability estimate $\{p_i\}_{i=1}^{n}$ of n samples in descending order. $Precision@n$ corresponds to recall of 1 because the classification threshold is set so small that all samples are predicted positive. Same as the previous section, we assign Bernoulli random variables $\{y_i\}_{i=1}^{n}$ as true labels of video samples. Under such circumstances, we have

$$PREC = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{19}$$

and the expected precision becomes

$$\widehat{PREC} = E_{\substack{y_i \sim Ber(p_i) \\ i=1,\ldots,n}} \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{1}{n} \sum_{i=1}^{n} p_i \tag{20}$$

Therefore, the approximation of video x_j can be viewed as

$$\phi_{PREC}(x_j) = \frac{1}{n-1}(\sum_{i=1}^{n} p_i - p_j) \tag{21}$$

Dataset	HMDB51						UCF50					
#L	Rand	Pmax	Entro	Usdm	Rank	USAP	Rand	Pmax	Entro	Usdm	Rank	USAP
10c	43.47	43.47	43.47	43.47	43.47	43.47	83.88	83.88	83.88	83.88	83.88	83.88
11c	44.17	44.38	44.39	**44.83**	44.43	44.60	83.87	84.95	85.60	84.75	**85.75**	85.24
12c	44.49	45.57	45.02	**45.89**	45.46	45.87	84.76	85.96	86.12	86.08	**86.15**	85.97
13c	44.78	46.79	46.17	**47.02**	46.43	46.87	85.01	87.45	86.57	86.12	**87.58**	87.79
14c	45.15	47.79	47.44	**47.94**	47.30	47.59	85.81	**87.81**	87.15	86.46	88.14	88.25
15c	45.59	**48.71**	48.18	48.60	48.01	48.41	86.62	87.89	**87.99**	86.84	88.65	88.83
16c	46.20	**49.44**	48.90	48.30	48.71	49.06	86.90	88.44	88.66	87.13	88.96	**89.28**
17c	46.47	**50.18**	49.66	48.67	49.31	49.93	87.49	89.09	88.92	87.38	89.16	**90.14**
18c	46.99	**50.88**	50.49	49.36	50.02	50.69	87.43	88.95	88.98	88.70	89.42	**89.94**
19c	47.39	51.36	51.02	50.00	50.71	**51.53**	87.33	89.14	89.25	89.27	89.57	**90.58**
20c	47.96	52.22	51.70	50.57	51.33	**52.94**	87.89	89.72	89.33	89.10	89.91	**90.82**
21c	48.26	52.83	52.59	51.02	51.99	**53.89**	88.30	89.97	89.40	90.23	89.89	**91.36**
22c	48.43	53.35	53.22	51.44	52.46	**55.44**	88.69	90.28	89.68	90.58	90.07	**91.29**
23c	48.82	53.89	53.65	52.21	52.96	**55.29**	88.75	90.55	90.22	90.56	90.04	**91.27**
24c	49.14	54.37	54.09	52.58	53.55	**55.71**	88.80	90.87	90.45	90.47	90.10	**91.42**
25c	49.65	54.85	54.40	52.79	54.04	**56.04**	89.17	91.46	90.59	90.79	90.15	**91.65**
26c	50.07	55.30	54.69	53.22	54.44	**56.32**	89.33	91.65	90.81	91.19	90.32	**91.88**
27c	50.49	55.45	55.02	53.43	54.98	**56.74**	89.28	91.55	91.15	91.50	90.64	**92.02**
28c	51.02	55.89	55.76	53.88	55.27	**57.25**	89.50	91.71	91.21	91.62	90.72	**92.34**
29c	51.34	56.26	56.10	54.49	55.59	**57.52**	89.64	91.86	91.34	92.05	90.99	**92.30**
30c	51.72	56.69	56.53	54.62	56.01	**57.83**	90.09	92.14	91.36	91.96	91.23	**92.35**

Table 1: MAP (in percentages) of all compared methods on dataset HMDB51 and UCF50. The method with the highest MAP is in boldface.

Given the above approximation, we calculate $\phi_{PREC}(\cdot)$ for each unlabeled videos and each one-vs-all classifier output, and sum $\phi_{PREC}(\cdot)$ over all classes. The videos with small sums are discarded for efficiency purposes. The screening rule runs in $\mathcal{O}(cn)$ for c-class classification with n unlabeled samples.

4 Experimental Results

Experiment Setting For each video clip, a level-three MIFS [Lan *et al.*, 2015] feature extractor is used to extract fixed-length feature, and Logistic Regression with parameter $C = 100$ is used as the underlying linear classifier to conduct one-vs-rest classification. Note that the purpose of this paper is not to obtain state-of-the-art action recognition results, but to demonstrate the necessity to consider underlying evaluation metric when building active learning algorithms for action recognition. For each dataset, we use the official training set as unlabeled set and official testing set as the unseen testing set. All active learning methods select data from the unlabeled dataset, and one-vs-rest classifiers are trained on all labeled data and tested on testing set. We use Mean Average Precision (MAP) as the evaluation metric. For a c-class video classification, we randomly select 10 videos of each class as initial labeled dataset. As a result, there are $10c$ labeled videos in the beginning. We iteratively select c videos until the labeling budget is reached. All experiments are repeated with different initial labeled data, and the averaged MAP is reported.

Compared methods We compare our *USAP* method with the following multi-class baselines:

- *Rand*: the method that selects unlabeled videos uniformly at random
- *Pmax*: the method that chooses uncertain samples based-on the maximum posterior probability in Eq. (3) [Lewis and Gale, 1994]
- *Entro*: the uncertainty sampling method based on Shannon entropy in Eq. (4)
- *Usdm*: the method that maximizes entropy along with diversity [Yang *et al.*, 2015]
- *Rank*: the method that combines entropy with Mutual Information as diversity [Chakraborty *et al.*, 2015]

Datasets Three representative datasets are used:

The *HMDB51* dataset [Kuehne *et al.*, 2011] has 51 action classes and 6766 video clips extracted from digitized movies and YouTube. [Kuehne *et al.*, 2011] provides both original videos and stabilized ones. Only original videos are used in this paper. There are three official train-test splits for this dataset. For each split, we run all active learning methods with two different initial labeled set.

The *Hollywood2* dataset [Marszalek *et al.*, 2009] contains 12 action classes and 1707 video clips that are collected from 69 different Hollywood movies. It has 1707 videos in total with a pre-defined split of 823 training videos and 884 test videos. We run each active learning method with five different initial labeled set on this dataset.

The *UCF50* dataset [Reddy and Shah, 2013] has 50 action classes spanning over 6618 YouTube videos clips that can be split into 25 groups. The video clips in the same group are generally very similar in background. We use one group as testing data and the other 24 groups as unlabeled data, so there are 25 different splits for this dataset. We use the first

Dataset	Hollywood2					
#L	Rand	Pmax	Entro	Usdm	Rank	USAP
10c	47.50	47.50	47.50	47.50	47.50	47.50
11c	48.27	48.54	48.92	**48.95**	48.09	48.53
12c	49.51	49.18	49.29	**49.59**	48.99	49.25
13c	50.06	49.85	50.06	**50.08**	49.37	49.98
14c	**50.96**	50.12	50.67	50.53	50.13	50.65
15c	**51.85**	50.73	50.88	50.41	50.80	**51.85**
16c	52.52	51.48	51.50	51.82	51.27	**52.86**
17c	52.99	52.11	52.27	52.53	52.17	**53.94**
18c	53.43	53.10	53.21	53.19	52.35	**54.56**
19c	53.99	54.01	53.56	53.54	52.95	**55.08**
20c	54.44	54.58	53.48	53.73	53.64	**55.94**
21c	54.76	54.87	54.14	53.66	54.34	**56.61**
22c	55.20	55.26	54.77	54.06	54.73	**56.66**
23c	55.51	55.77	55.32	55.34	55.21	**57.22**
24c	55.89	56.47	56.05	55.67	55.84	**58.11**
25c	56.45	57.02	56.16	56.10	56.13	**58.23**
26c	57.08	57.50	56.55	56.27	56.71	**58.37**
27c	57.52	57.63	57.20	56.83	56.92	**58.71**
28c	57.77	58.01	57.45	57.43	57.24	**59.17**
29c	58.28	58.25	57.84	58.05	57.73	**59.31**
30c	58.74	58.25	58.26	58.15	58.27	**59.61**

Table 2: MAP (in percentages) of all compared methods on dataset Hollywood2. The method with the highest MAP is in boldface.

5 of them to test all active learning algorithms, each with one initial labeled set.

To speed up our algorithm, on large datasets *UCF50* and *Hmdb51*, we use the fast screen rule in Section 3.4 to filter out uninformative video samples so that the size of the unlabeled pool does not exceed 2000. Algorithm 1 is then used to select uncertain samples accordingly on the remaining videos.

Results We illustrate the performance of the active learning algorithms for action recognition application in Table 1 and 2. In the two tables, each row corresponds to the number of selected videos, and each column corresponds to an active learning algorithm.

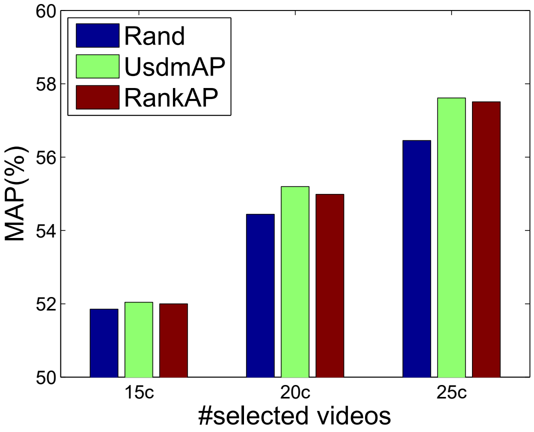

Figure 1: MAP (in percentages) of Rand, UsdmAP and RankAP on dataset *Hollywood2*

As shown in the two Tables, we have three observations. First, on dataset *HMDB51* and *UCF50*, all active learning algorithms generally outperform random sampling, while on *hollywood2*, all other active learning algorithms (except ours) fail to outperform random sampling. This is because movie videos of this dataset sometimes have more than one labels, which misleads the conventional uncertainties. Second, our algorithm outperforms the other active learning algorithms when the number labeled samples becomes large. This shows that optimizing AP for active learning is effective. Note that *Usdm* and *Rank* sometimes have higher MAP than our algorithm in the initial stage of active learning, which is the effect of the diversity used. Third, our algorithm selects fewer video samples than other active learning algorithms to achieve the same MAP when the labeled set becomes large. For example in *hollywood2*, our algorithm requires $24c$ labeled videos to achieve 58% MAP while *Pmax* and *Entro* need $28c$ and $30c$ respectively.

Combine USAP with diversity In order to integrate our AP-based uncertainty with diversity, we replace entropy-based uncertainty with our AP-based uncertainty (ϕ in Algorithm 1) for algorithm *Rank* and *Usdm*. The two new algorithms, denoted as *RankAP* and *UsdmAP*, are tested on dataset *Hollywood2* with $15c$, $20c$, and $25c$ labeled videos.

In Figure 1, *UsdmAP* and *RankAP* outperforms *Rand* in three different number of labeled videos, namely $15c$, $20c$ and $25c$, while Table 2 shows that, *without* our uncertainty measure, *Usdm* and *Rank* sometimes has slightly lower MAP than *Rand*. This result demonstrates that under MAP, our *USAP* uncertainty measure is able to find more informative videos than the conventional Shanon entropy, combined with existing diversity measures in *Usdm* and *Rank*.

5 Conclusion and Future Work

In this paper, we propose a novel active learning algorithm for action recognition by maximizing expected AP, and treating the unknown *true labels* of unlabeled videos as independent Bernoulli random variables. The expected AP can be formalized and calculated on the joint distribution of all random variables. Then the uncertainty of each unlabeled video is defined as the expected AP of all other unlabeled videos. The proposed uncertainty sampling algorithm is tested on three real-word action recognition datasets. Experiments show that our algorithm outperforms other uncertainty-based algorithms. As future work, we plan to explore new active learning methods under other metrics such as mean accuracy, F-measure and AUC.

Acknowledgments

This work is supported in part by China National 973 program 2014CB340301. We also acknowledge the Data to Decisions and the Cooperative Research Centers program (D2D CRC) for funding Yi Yang's research.

References

[Behl *et al.*, 2014] Aseem Behl, CV Jawahar, and M Pawan Kumar. Optimizing average precision using weakly supervised data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1011–1018, 2014.

[Chakraborty *et al.*, 2015] Shayok Chakraborty, Vineeth Balasubramanian, Qian Sun, Sethuraman Panchanathan, and Jieping Ye. Active batch selection via convex relaxations with guaranteed solution bounds. *IEEE transactions on pattern analysis and machine intelligence*, 37(10):1945–1958, 2015.

[Culver *et al.*, 2006] Matt Culver, Deng Kun, and Stephen Scott. Active learning to maximize area under the roc curve. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*, pages 149–158. IEEE, 2006.

[Dutt Jain and Grauman, 2016] Suyog Dutt Jain and Kristen Grauman. Active image segmentation propagation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2864–2873, 2016.

[Kuehne *et al.*, 2011] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2556–2563. IEEE, 2011.

[Lan *et al.*, 2015] Zhengzhong Lan, Ming Lin, Xuanchong Li, Alex G Hauptmann, and Bhiksha Raj. Beyond gaussian pyramid: Multi-skip feature stacking for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 204–212, 2015.

[Lewis and Gale, 1994] David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12. Springer-Verlag New York, Inc., 1994.

[Liu and Ferrari, 2017] Buyu Liu and Vittorio Ferrari. Active learning for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4363–4372, 2017.

[Long *et al.*, 2010] Bo Long, Olivier Chapelle, Ya Zhang, Yi Chang, Zhaohui Zheng, and Belle Tseng. Active learning for ranking through expected loss optimization. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 267–274. ACM, 2010.

[Marszalek *et al.*, 2009] Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2929–2936. IEEE, 2009.

[Perronnin *et al.*, 2010] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. *Computer Vision–ECCV 2010*, pages 143–156, 2010.

[Ramirez-Loaiza *et al.*, 2017] Maria E Ramirez-Loaiza, Manali Sharma, Geet Kumar, and Mustafa Bilgic. Active learning: an empirical study of common baselines. *Data Mining and Knowledge Discovery*, 31(2):287–313, 2017.

[Reddy and Shah, 2013] Kishore K Reddy and Mubarak Shah. Recognizing 50 human action categories of web videos. *Machine Vision and Applications*, 24(5):971–981, 2013.

[Roy and McCallum, 2001] Nicholas Roy and Andrew McCallum. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, pages 441–448, 2001.

[Settles, 2010] Burr Settles. Active learning literature survey. *University of Wisconsin, Madison*, 52(55-66):11, 2010.

[Seung *et al.*, 1992] H Sebastian Seung, Manfred Opper, and Haim Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294. ACM, 1992.

[Sharma and Bilgic, 2017] Manali Sharma and Mustafa Bilgic. Evidence-based uncertainty sampling for active learning. *Data Mining and Knowledge Discovery*, 31(1):164–202, 2017.

[Triantafillou *et al.*, 2017] Eleni Triantafillou, Richard Zemel, and Raquel Urtasun. Few-shot learning through an information retrieval lens. In *Advances in Neural Information Processing Systems*, pages 2252–2262, 2017.

[Wang and Schmid, 2013] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013.

[Wang *et al.*, 2011] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176. IEEE, 2011.

[Yang *et al.*, 2015] Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander G Hauptmann. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, 113(2):113–127, 2015.

[Yu *et al.*, 2006] Kai Yu, Jinbo Bi, and Volker Tresp. Active learning via transductive experimental design. In *Proceedings of the 23rd international conference on Machine learning*, pages 1081–1088. ACM, 2006.

[Yue *et al.*, 2007] Yisong Yue, Thomas Finley, Filip Radlinski, and Thorsten Joachims. A support vector method for optimizing average precision. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 271–278. ACM, 2007.

[Zhou *et al.*, 2017] Zongwei Zhou, Jae Shin, Lei Zhang, Suryakanth Gurudu, Michael Gotway, and Jianming Liang. Fine-tuning convolutional neural networks for biomedical image analysis: actively and incrementally. In *IEEE conference on computer vision and pattern recognition, Hawaii*, pages 7340–7349, 2017.