ABSTRACT
Cross-domain video concept detection is a challenging task due to the distribution difference between the source domain and target domain. In order to avoid expensive labeling the target-domain data, Active Learning can be used to incrementally learn a target classifier by reusing the one in the source domain. It uses a discriminative query strategy and picks the most ambiguous samples to label, which could fail if the distribution difference is too large. In this paper, to deal with large difference in data distributions, we propose a generative query strategy which is then combined with the existing discriminative one to yield a hybrid method. This method adaptively fits the distribution differences and gives a mixture strategy that performs more robustly compared to both single strategies. Experimental results on TRECVID semantic concept detection task demonstrate superior performance of our hybrid method.

Categories and Subject Descriptors
H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing

General Terms
Algorithms, Experimentation, Performance

Keywords
Active Learning, Cross-Domain Video Concept Detection

1. INTRODUCTION
Video concept detection aims to determine the presence or absence of a semantic concept in video shots, which serves as an important intermediate step in concept-based video retrieval. Many automatic concept detection techniques have been developed, among which the most popular one is to use multiple binary classifiers to predict the relevance between video shots and a given concept.

One challenge of these detection approaches is that multimedia data come from a variety of data domains (e.g., news, documentary, entertainment) and a classifier trained on one domain may perform poorly on another [1]. The reason is that the data from two different domains may have different distributions, though they still share the same concepts. If we always train a new classifier for every new domain, we face expensive labeling efforts to provide sufficient training samples. It’s better to reuse the existing classifier in the source domain.

Transfer learning is a major approach for cross-domain problems, which aims to automatically transfer the knowledge learnt from a source domain to a related but different target domain [4]. Yang et al. [1] adapt the source domain classifier to the target domain by learning a delta function between the decision functions in the source and target domains. Their method seeks a new decision boundary that is close to the boundary in the source domain and also separates the labeled examples in target dataset well. Jiang et al. [5] train a SVM classifier using both the source and target data, with the target data weighted by the distance to the support vectors of the source data. Duan et al. [14] learn a kernel function and a SVM classifier simultaneously by minimizing the risk of both SVM and the distribution mismatch of the source and target datasets. The drawback of these transfer learning methods is that they rely on strong assumptions that might not be held in real data, as pointed by previous research [6]. For example, when the distributions in the two domains are quite different, Yang’s method [1] fails because the two decision boundaries may differ significantly. Our work approaches the cross-domain video concept detection problem from a different view. Instead of automatic transfer learning without human labeling, we query for labels from a small subset of informative samples.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Figure 1. Illustration of discriminative and generative query strategies. Squares and circles represent different classes; blue and red denote the source and target domains, respectively. When the distributions of the two domains are quite different, the samples selected by the discriminative query strategy (contained in black dash) are less informative than those selected by the generative query strategy (contained in green dash) which selects samples with low likelihood given the source-domain distribution.

The principle way for selective labeling is called Active Learning [2], which has been extensively studied in the last two decades. A
number of query strategies have been proposed, such as query by uncertainty [7], query by committee [8], and query by expected model change [9]. Many active learning methods choose samples close to the decision boundary [12], since these samples are most ambiguous for the current classifier. However, the assumption of active learning is that the unlabeled data have same distributional properties as the original dataset. In other words, both source and target data are from the same domain. If their distributions are different, the samples close to the boundary may not be informative in building the new classifier.

In the context of cross-domain video concept detection, the traditional query strategy, i.e., querying samples close to the discriminative boundary, may still be applicable, if the data distributions in the source and target domain are close. This query strategy can be regarded as a discriminative strategy, because it queries the new data from a discriminative view. However in the case that the distributions of the source and target domain differ significantly, a generative query strategy, which queries the samples unlikely to be generated by the source distribution, is likely to be more effective. This query strategy essentially accounts for distributional differences, and explores regions where the old decision boundary is never relevant (see Figure 1).

In this paper, we propose a novel cross-domain active learning approach – Hybrid Active Learning, which combines two different query strategies: one is discriminative, which selects samples close to the current discriminative boundary; the other is generative, which selects samples unlikely to be generated from current distribution. Additionally, we present an adaptive way to combine the two query strategies, leading to excellent performance regardless of how much different the distributions of the two domains are. We evaluate the proposed method on cross-domain video concept detection on TRECVID collections of 2005 and 2007 [3], which include a variety of domain data ranging from news videos to documentary videos. Experimental results show that our Hybrid Active Learning approach performs better than classifiers trained on the source domain only, as well as active learning using either a discriminative or a generative strategy alone.

2. HYBRID ACTIVE LEARNING

An active learning method often runs in several rounds. It starts with a model trained from a labeled training set. Then in each round, it uses a query strategy to rank all unlabeled samples according to their impact on the learning task, and picks a small set of most informative samples for query. After that, it adds the queried samples as well as their labels to the training set, and re-trains the model. The advantage of active learning methods is that they aim at querying most informative samples, which helps to significantly reduce the human labeling cost. In this section, we first introduce two different query strategies, namely discriminative active learning and generative active learning. Then we present the proposed Hybrid Active Learning approach.

2.1 Discriminative Active Learning (Dis-AL)

For video concept detection, following previous studies [1, 5] we use Support Vector Machine (SVM) as our classifier. Basically, given a set of samples $\{x\}$ and their corresponding labels $\{y\}$, SVM seeks a hyperplane $f(x) = w \cdot \phi(x) + b$, which separates the samples into different classes with a maximum margin criteria [10, 11]. Here $w$ and $b$ are model parameters. $\phi(x)$ is the feature mapping function which projects raw features to a high dimensional feature space. The discriminative query strategy for SVM is to select samples close to the boundary. Intuitively, the samples close to the decision boundary are the most ambiguous ones to predict labels by the current classifier. This strategy has been explored in a number of previous studies, and has been theoretically proved to efficiently reduce the hypothesis space [12]. Thus, samples having the shortest distances to the hyperplane are selected. Mathematically, Dis-AL queries the samples using the criteria below [12]:

$$x^* = \min_{x_i} |w \cdot \phi(x_i) + b|$$

where $|w \cdot \phi(x_i) + b|$ is proportional to the distance between a sample and the hyperplane. Such a strategy can be expected to work when the source and target data have similar distributions.

2.2 Generative Active Learning (Gen-AL)

When the distribution of the target data is quite different from the source, the discriminative boundary trained from the source data is not reliable for the target data. So we want to use the distribution information of the source data to find samples most unlikely to be generated in the source domain.

For simplicity, we assume positive samples and negative samples of the source data following two Gaussian distributions in the kernel space. The likelihood of one sample $x_i$ given a distribution $D$ (can be either the distribution of positive samples, $D^+$ or the distribution of negative ones, $D^-$) is computed as follows:

$$p_D(x_i) = \frac{1}{(2\pi)^d/2|\Sigma|^{1/2}}\exp\left[-\frac{1}{2}((\phi(x_i) - \mu)^\Sigma^{-1}(\phi(x_i) - \mu)\right]$$

where $d$ is the dimensionality of each sample, $\mu$ is the mean of the Gaussian distribution in the kernel space and $\Sigma$ is the corresponding covariance matrix.

Gen-AL tends to query samples in the target domain unlikely to be covered by either the positive or negative distribution of the source domain, i.e. with low likelihood in either distribution, thus their labels are highly informative for the classifier. In each round, Gen-AL first computes the maximum likelihood of each sample belonging to either distribution:

$$p_{m_i} = \max(p_{D^+}(x_i), p_{D^-}(x_i))$$

Here $p_{D^+}(x_i)$ and $p_{D^-}(x_i)$ denote the likelihood of $x_i$ given $D^+$ and $D^-$, respectively. Then, it sorts all samples according to their $p_m$ in ascending order. After that, Gen-AL picks the top $k$ samples for query.

Note that in the kernel space, the distance of a sample $\phi(x_i)$ to the mean of a distribution $\mu$ can be calculated as follows:

$$\|(\phi(x_i) - \mu)\|^2$$

$$= \|(\phi(x_i) - \frac{1}{n}\sum_{j=1}^n \phi(x_j))\|^2$$

$$= k(x_i, x_i) - \frac{2}{n}\sum_{j=1}^n k(x_i, x_j) + \frac{1}{n^2}\sum_{j=1}^n \sum_{j=1}^n k(x_j, x_j)$$

where $k(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$ and $n$ is the number of samples of the source data. In our implementation, we simply treat $\Sigma$ of the positive and negative distributions to be equal. As a result, the decision boundary between the positive and negative distributions is linear, which coincides with SVM.

2.3 Hybrid Active Learning (Hyb-AL)

As mentioned above, a discriminative query strategy is suitable for the case where the source and target domains have similar
distributions, while a generative query strategy is suitable when the two domains have significantly different distributions. Since in practice we do not know how much the two distributions differ, it’s beneficial to use a hybrid query strategy that combines the advantages of both discriminative and generative ones.

Algorithm 1 Hybrid Active Learning Method

Inputs:
- $S$: source data (fully labeled);
- $T_1$: labeled data in target dataset $T$;
- $T_u$: unlabeled data in target dataset $T$;
- $S_{Dis}$: discriminative query strategy;
- $S_{Gen}$: generative query strategy;
- $K$: number of rounds; $N$: total number of queries in each round;
- $w_d$: weight of $S_{Dis}$; $w_g$: weight of $S_{Gen}$;
- $\eta$: scaling factor for adapting the weights of two strategies;
- $N_{wd}$: total number of queries in each round;

Outputs: model trained by using data from $\{S \cup T_1\}$.

Procedure:
1. Query samples by the two strategies:
   Query $N w_d(i)$ samples from $T_u$ using $S_{Dis}$; $n_d(i)$ is the set of samples it queries, $|n_d(i)|_+$ is the ratio of positive samples in $n_d(i)$;
   Query $N w_g(i)$ samples from $T_u - n_d(i)$ using $S_{Gen}$; $n_g(i)$ is the set of samples it queries, $|n_g(i)|_+$ is the ratio of positive samples in $n_g(i)$.

2. Update the dataset and the training model:
   Remove $\{n_d(i) \cup n_g(i)\}$ from $T_u$, and add them to $T_1$;
   Train a SVM model by using labeled dataset $\{S \cup T_1\}$.

3. Update the weights:
   if $|n_d(i)|_+ > |n_g(i)|_+$ then
   \[ w_d(i+1) = \max(\eta w_d(i), 1), \]
   \[ w_g(i+1) = 1 - w_d(i+1); \]
   else if $|n_d(i)|_+ < |n_g(i)|_+$ then
   \[ w_g(i+1) = \max(\eta w_g(i), 1), \]
   \[ w_d(i+1) = 1 - w_g(i+1); \]
   else $|n_g(i)|_+ = |n_d(i)|_+$ then
   \[ w_g(i+1) = w_g(i), \]
   \[ w_d(i+1) = w_d(i); \]
   end

End

The idea of our hybrid method is simple: In each round, the total number of samples to be queried is fixed. Each strategy (i.e. discriminative or generative) has its own weight, which determines the portion of samples it can query. If one strategy has a higher weight than the other, it is allowed to query more samples. At the end of each round, the weights of different strategies are adapted according to their effectiveness in exploring informative samples in the target domain. In our implementation, the effectiveness of a strategy is measured by the ratio of positive samples in the sample set it queries. The intuition is that in video concept detection, positive samples usually carry more information, because most datasets are unbalanced. Initially, we set the weights of the two strategies to be equal. The proposed method is shown in Algorithm 1.

3. EXPERIMENTAL RESULTS

In this work, we investigate the cross-domain video concept detection problem using two TRECVID video benchmark datasets [3]. The development set of TRECVID 2005 is used as the source dataset which contains 61,901 shots extracted from 108 hours of broadcast news videos from international programs. The TRECVID 2007 dataset is used as target dataset, which contains 21,532 shots extracted from 60 hours of news magazines, documentaries, science news, and educational programming videos. Similar to [5], we use 17,520 keyframes as the training set and 2,012 keyframes as the evaluation set. TRECVID 2005 and TRECVID 2007 are from different domains which have different programs, but they contain similar concepts. 36 concepts defined by LS.COM-Lite lexicon [13] have been manually annotated in both of these two datasets. Each shot is represented by one keyframe, and is described by a 346-dimensional feature vector consisting of three standard low-level visual features. They are a 225-dimensional grid-color moment feature, a 48-dimensional Gabor texture, and a 73-dimensional edge direction histogram.

To ensure different query strategies are comparable, we use the same RBF kernel function $k(x_i, x_j) = \exp(-\gamma ||x_i - x_j||^2)$ with $\gamma = 0.25$ in all methods. Also we set the penalty term $C = 10$ for training all SVMs. For evaluation, the classifier output is converted to a descending rank list based on the probability score and average precision (AP) is calculated over all relevant examples in the rank list, which is a standard metric in TRECVID. Mean average precision (MAP) is used to average all the APs of 36 concepts for evaluation.

Query Strategy: we examine four different query strategies. For each strategy, the top 100 samples are selected based on each query criteria. After querying, labels for samples are added to the source dataset to re-train the classifier and we evaluate the performance on the evaluation set. 10 rounds are performed for each query strategy. Four different strategies for comparison are listed below:

- **Dis-AL**: query samples which are close to the discriminative boundary.
- **Gen-AL**: query samples based on their likelihood to positive and negative sets of the source data, selecting samples with low likelihood to both.
- **Hyb-AL**: query samples based on Hybrid Active Learning method which combines Dis-AL and Gen-AL.
- **Random**: query samples randomly in each round. This is the baseline method.

Figure 2 shows the performance of the final round of four query strategies in terms of AP for all 36 concepts. We can see:

(a) For 8 concepts, Dis-AL performs best. A reasonable explanation is that the source and target datasets have similar distributions in these concepts, thus the trained decision boundary on the source domain is still reliable;
(b) For 7 concepts, Gen-AL works better than other methods, probably due to significant difference in the source and target distributions. Consequently, the old decision boundary works poorly but the generative query becomes more effective;
(c) For most of the remaining concepts, Dis-AL and Gen-AL perform comparatively and are both better than Random;
However, for all concepts, Hyb-AL is comparable to or better than the best of Dis-AL and Gen-AL. Meanwhile Hyb-AL is significantly better than the worst of them. This is because it adaptively finds the best query strategy, thus is more robust to different cases.

Figure 3 shows the evolution in MAP of all 36 concepts according to different query strategies. The results show that:

(a) All active learning methods outperform the classifier trained on the source-domain data only. This confirms that active learning is a promising approach for cross-domain video concept detection;
(b) Hyb-AL performs slightly worse than Dis-AL in the first several rounds, but soon outperforms all other query strategies in later rounds. This demonstrates that Hyb-AL is capable of finding the best strategy in an adaptive manner.

4. CONCLUSION
In this paper, we propose Hybrid Active Learning, a novel active learning method which is capable of working on two different domains. The proposed method adaptively combines both discriminative and generative query strategies. Experimental results show that Hybrid Active Learning is a promising approach for cross-domain concept detection. Our future work will develop a better generative query strategy by also considering the variance of the distributions. Also, we will further explore active learning with multiple query strategies.

5. REFERENCES