Columbia University TRECVID2007 High-Level Feature Extraction

Shih-Fu Chang, Wei Jiang, Akira Yanagawa, Eric Zavesky
{sfchang,wjiang,akira,emz}@ee.columbia.edu
Columbia University
Department of Electrical Engineering
http://www.ee.columbia.edu/dvmm

Draft, October 22, 2007

1 Description of Submissions

High-level feature extraction

- **A_COL_base6_T2_6** (R6): train on TRECVID2007 development data, multi-parameter models with average fusion across several modalities (referred to as *target baseline*)
- **A_COL_xd5_T5_5** (R5): train on TRECVID2007 development data and TRECVID2005 development data using feature replication, multi-parameter models with average fusion across several modalities (referred to as *xd* or *replication*)
- **A_COL_bcrf_base4_T7_4** (R4): trained with contextual scores from TRECVID2005 based models (using *columbia374*) and with TRECVID2007 development (referred to as *BCRF*) then average-fused with *target baseline*
- **A_COL_bcrf_xd_base3_T14_3** (R3): average fusion of *BCRF*, *target baseline*, and *xd replication*
- **A_COL_bcrf_xd_base_col3742_T16_2** (R2): average fusion of *BCRF*, *target baseline*, *xd replication*, and *columbia374*
- **A_COL_best_of_all1_T17_1** (R1): choose best-performing classifier for each concept over a validation data subset from all above submissions
- **col374**: not submitted, but publicly available model (see [16]) set covering 374 concepts in the LSCOM ontology and trained on only 60% of the TRECVID2005 development data (referred to as *source models*)

Abstract

One difficulty in the HLF task this year was changing the applied domain from news video to foreign documentary videos. Classifiers trained in prior years performed poorly if naively applied, and classifiers trained on the 2007 data alone may suffer from too few positive training samples. This year we address this new fundamental problem how to efficiently and effectively adapt models learned from an old domain to a significantly different one. Investigation of this topic complements very well the scalability issue discussed in TRECVID 2006 how to leverage the resource of a large concept detector pool (e.g., Columbia 374) to improve accuracy of individual detectors.

We developed and tested a new cross-domain SVM (CDSVM) algorithm for adapting previously learned support vectors from one domain to help classification in another domain. Performance gain is obtained with almost no additional computational cost. Also, we conduct a comprehensive comparative study of the state-of-the-art SVM-based cross-domain learning methods.

To further understand the underlying contributing factors, we propose an intuitive selection criterion to determine which cross-domain learning method to use for each concept. Such a prediction mechanism is important since there are a multitude of promising methods for adapting old models to new domains, and thus judicious selection is a key to applying the right method under the right context (e.g., size of training data in new/old domains, variation of content between two domains, etc). Although there is no single method that universally outperforms other options, with adequate prediction mechanisms, we will be able to apply the right adaptation approach in different conditions, and demonstrate 22% performance improvement for mid-frequency or rare concepts.

2 Introduction

There is a common issue for machine learning problems: the amount of available test data is large and growing, but the amount of labeled data is often fixed and quite small. Video data, labeled for semantic concept classification is no exception. For example, in high-level concept classification tasks (TRECVID [14]), new corpora may be added annually from unseen sources like foreign news channels or audio-visual archives. Ideally, one desires the same low error rates when reapplying models derived from previous source domain $\mathcal{D}^s$ to a new, unseen target domain $\mathcal{D}^t$, often referred to as domain adaptation or cross-domain learning. Recently several different approaches has been proposed toward this direction in the machine learning society [4, 5, 17]. The high-level feature extraction task of TRECVID2007 provides a large amount of cross-domain data sets for evaluating and comparing these methods. In TRECVID2007, we try to tackle this challenging issue and make contributions in two folds. First, a new *Cross-Domain SVM* (*CDSVM*) algorithm is developed for adapting previously learned support vectors from source $\mathcal{D}^s$ to help detect concepts in target $\mathcal{D}^t$. Better precision can be obtained with almost no additional computational cost. Second, a comprehensive summary and comparative study of the state-of-the-art SVM-based cross-domain learning algorithms is given. By treating the TRECVID2007 data set as the target domain $\mathcal{D}^t$ and treating the TRECVID2005 data set as the source domain $\mathcal{D}^s$, these algorithms are evaluated over the latest large-scale TRECVID benchmark data. Finally, a simple but effective criterion is proposed to determine if and which cross-domain method should be used.

The rest of this paper is organized as follows. Section 3 gives an overview of many state-of-the-art SVM-based cross-domain learning methods, ordered in decreasing computational cost. Section 3.3.3 introduces our CDSVM algorithm. We also review the BCRF approach which explores the inter-concept relations. In section 4 we discuss our submissions for TRECVID2007 high-level feature extraction task and in section 5 we compare the performance of many cross-domain learning algorithms. Finally, in section 6 we provide experimental conclusions and next steps for research.

3 Approach Overview

The cross-domain learning problem can be summarized as follows. Let $\mathcal{D}^t$ denote the *target data set*, which consists of two subsets: the labeled subset $\mathcal{D}_l^t$ and the unlabeled subset $\mathcal{D}_u^t$. Let $(\mathbf{x}_i, y_i)$ denote a data point where $\mathbf{x}_i$ is a d dimensional feature vector and y_i is the corresponding class label. In this work we only look at the binary classification problem, i.e., $y_i = \{+1, -1\}$. In addition to $\mathcal{D}^t$, we have a *source data set* $\mathcal{D}^s$ whose distribution is different from but related to that of $\mathcal{D}^t$. A binary classifier $f^s(\mathbf{x})$ has already been trained over this source data set $\mathcal{D}^s$. Our goal is to learn a classifier $f(\mathbf{x})$ to classify the unlabeled target subset $\mathcal{D}_u^t$.

As $\mathcal{D}^t$ and $\mathcal{D}^s$ have different distributions, $f^s(\mathbf{x})$ will not perform well for classifying $\mathcal{D}_u^t$. Conversely, we can train a new classifier $f^t(\mathbf{x})$ based on $\mathcal{D}_l^t$ alone, but when the number of training samples $|\mathcal{D}_l^t|$ is small, $f^t(\mathbf{x})$ may not give robust performance. Since $\mathcal{D}^s$ is related to $\mathcal{D}^t$, utilizing information from source $\mathcal{D}^s$ to help classify target $\mathcal{D}_u^t$ should yield better performance. This is fundamental the motivation of cross-domain learning. In this section, we briefly summarize and discuss many state-of-the-art SVM-based cross-domain learning algorithms.

3.1 Standard SVM Applied in New Domain

Without cross-domain learning, the standard *Support Vector Machine* (*SVM*) [15] classifier can be learned based on the labeled subset $\mathcal{D}_l^t$ to classify the unlabeled set $\mathcal{D}_u^t$. Given a data vector $\mathbf{x}$, SVMs determine the corresponding label by the sign of a linear decision function $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$. For learning non-linear classification boundaries, a kernel mapping ϕ is introduced to project data vector $\mathbf{x}$ into a high-dimensional feature space $\phi(\mathbf{x})$, and the corresponding class label is now given by the sign of $f(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x}) + b$. The primary goal of an SVM is to find an optimal separating hyperplane that gives a low generalization error while separating the positive and negative training samples. This hyperplane is determined by giving the largest margin of separation between different classes, i.e. by solving the following problem:

$$\begin{aligned} &\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||_2^2 + C\sum\nolimits_{i=1}^{N_l^t} \epsilon_i \\ &s.t.\ \ y_i\mathbf{w}^T\phi(\mathbf{x}_i) + b \geq 1 - \epsilon_i,\ \ \epsilon_i \geq 0,\ \ \forall(\mathbf{x}_i, y_i) \in \mathcal{D}_l^t \end{aligned} \tag{1}$$

where ϵ_i is the penalizing variable added to each data vector $\mathbf{x}_i$; C determines how much error an SVM can tolerate. One very simple way to perform cross-domain learning is to learn new models over all possible samples, called *Combined SVM* in this paper. The primary motivation for this method is that when the size of data in target domain is small, the target model will benefit from a high count of training samples present in $\mathcal{D}^s$ and should therefore be much more stable than a model trained on $\mathcal{D}^t$ alone. However, there is a large time cost for learning with this method due to the increased number of training samples from $|\mathcal{D}^t|$ to $|\mathcal{D}^s| + |\mathcal{D}^t|$.

3.2 Transductive Localized SVM (LSVM)

To decrease generalization error in classifying unseen data $\mathcal{D}_u^t$ in the target domain, transductive SVM methods [5, 9] incorporate knowledge about the new test data into the SVM optimization process so that the learned SVM can accurately classify test data.

The *Localized SVM* (*LSVM*) tries to learn one classifier for each test sample based on its local neighborhood. Given a test data vector $\hat{\mathbf{x}}_j$, we find its neighborhood in the labeled training set $\mathcal{D}_l^t$ based on similarity $\sigma(\hat{\mathbf{x}}_j, \mathbf{x}_i),\ \mathbf{x}_i \in \mathcal{D}_l^t$: $\sigma(\hat{\mathbf{x}}_j, \mathbf{x}_i) = \exp\left(-\beta||\hat{\mathbf{x}}_j - \mathbf{x}_i||_2^2\right)$. β controls the size of the neighborhood, i.e. the larger the β, the less influence each distant data point has. An optimal local hyper-plane is learned from test data neighborhoods by optimizing the following function:

$$\min_{\mathbf{w}} \frac{1}{2}||\mathbf{w}||_2^2 + C\sum_{i=1}^{N_l^t} \sigma(\hat{\mathbf{x}}_j, \mathbf{x}_i)\epsilon_i \quad (2)$$
$$s.t.\ \ y_i\mathbf{w}^T\phi(\mathbf{x}_i)+b \geq 1-\epsilon_i,\ \ \epsilon_i \geq 0,\ \ \forall(\mathbf{x}_i, y_i) \in \mathcal{D}_l^t$$

As the result, the classification of a test sample only depends on the support vectors in its local neighborhood.

Transductive SVM approaches can be directly used for cross-domain learning by using $\mathcal{D}_l^t \cup \mathcal{D}^s$ to take the place of $\mathcal{D}_l^t$ in Eqn.(2). Their major drawback is the computational cost, especially for large-scale data sets.

3.3 Cross-domain Adaptation Approaches

In the cross-domain learning problem, the source data set $\mathcal{D}^s$ and the target data set $\mathcal{D}^t$ are highly related. The following cross-domain adaptation approaches investigate how to use source data to help classify target data.

3.3.1 Feature Replication

Feature replication combines all samples from both $\mathcal{D}^s$ and $\mathcal{D}^t$, and tries to learn *generalities* between the two data sets by replicating parts of the original feature vector, $\mathbf{x}_i$ for different domains. This method has been shown effective for text document classification over multiple domains [8]. Specifically, we first zero-pad the dimensionality of $\mathbf{x}_i$ from d to $d(N-1)$ where N is the total number of adaptation domains, and in our experiments $N=2$ (one source and one target). Next we transform all samples from all domains as:

$$\hat{\mathbf{x}}_i^s = \begin{bmatrix} \mathbf{x}_i \\ \mathbf{0} \\ \mathbf{x}_i \end{bmatrix},\ \mathbf{x}_i \in \mathcal{D}^s \qquad \hat{\mathbf{x}}_i^t = \begin{bmatrix} \mathbf{x}_i \\ \mathbf{x}_i \\ \mathbf{0} \end{bmatrix}, \mathbf{x}_i \in \mathcal{D}^t$$

During learning, a model will be constructed that takes advantage of all possible training samples. Alike the combined method in section 3.1, this is most helpful when $\mathcal{D}^s$ can provide missing data for $\mathcal{D}^t$. However, unlike the combined method, learned SVs from the same domain as a test unlabeled sample (source-source or target-target) are given more preference by the the kernelized function of $\phi(\hat{\mathbf{x}}^s, \hat{\mathbf{x}}^s)$ or $\phi(\hat{\mathbf{x}}^t, \hat{\mathbf{x}}^t)$ compared to $\phi(\hat{\mathbf{x}}^t, \hat{\mathbf{x}}^s)$ because of the zero-padding operation. Unfortunately, due to the increase in dimensionality, there is also a large increase in model complexity and computation time during learning and evaluation of replication models.

3.3.2 Adaptive SVM

In [17], the *Adaptive SVM* (*A-SVM*) approach tries to adapt the a classifier $f^s(\mathbf{x})$, learned from $\mathcal{D}^s$ to classify the unseen target data set $\mathcal{D}_u^t$. In this approach, the final discriminant function is the average of $f^s(x)$ and the new "delta function" $\triangle f(\mathbf{x}) = \mathbf{w}^T\phi(\mathbf{x})+b$ learned from target set $\mathcal{D}_l^t$, i.e.,

$$f(\mathbf{x}) = f^s(\mathbf{x}) + \mathbf{w}^T\phi(\mathbf{x}) + b \quad (3)$$

where $\triangle f(\mathbf{x})$ aims at complementing $f^s(\mathbf{x})$ based on target $\mathcal{D}_l^t$. The basic idea of A-SVM is to learn a new decision boundary that is close to the original decision boundary (given by $f^s(\mathbf{x})$) as well as separating the target data.

One potential problem with this approach is the constraint that the new decision boundary should not be deviated far from the source classifier. This is generally a reasonable assumption when $\mathcal{D}^t$ is only incremental data for $\mathcal{D}^s$, i.e. $\mathcal{D}^t$ has similar distribution with $\mathcal{D}^s$. When $\mathcal{D}^t$ has a different distribution but comparable size than $\mathcal{D}^s$, such regularization constraint is problematic.

3.3.3 Cross-Domain SVM

In a recent work [10], we proposed a new method called *Cross-Domain SVM* (*CDSVM*). Our goal is to learn a new decision boundary based on the target data set $\mathcal{D}_l^t$ which can separate the unknown data set $\mathcal{D}_u^t$, with the help of $\mathcal{D}^s$. Let $\mathcal{V}^s = \{(\mathbf{v}_1^s, y_1^s), \ldots, (\mathbf{v}_M^s, y_M^s)\}$ denote the support vectors which determine the decision boundary and $f^s(\mathbf{x})$ be the discriminant function already learned from the source domain. Learned support vectors carry all the information about $f^s(\mathbf{x})$; if we can correctly classify these support vectors, we can correctly classify the remaining samples from $\mathcal{D}^s$ except for some misclassified training samples. Thus our goal is simplified and analogous to learning an optimal decision boundary based on the target data set $\mathcal{D}_l^t$ which can separate the unknown data set $\mathcal{D}_u^t$ with the help of $\mathcal{V}^s$.

Similar to the idea of LSVM, the impact of source data $\mathcal{V}^s$ can be constrained by neighborhoods. The rationale behind this constraint is that if a support vector $\mathbf{v}_i^s$ falls in the neighborhood of target data $\mathcal{D}^t$, it tends to have a distribution similar to $\mathcal{D}^t$ and can be used to help classify $\mathcal{D}^t$. Thus the new learned decision boundary needs to take into consideration the classification of this support vector. Let $\sigma(\mathbf{v}_j^s, \mathcal{D}_l^t)$ denote the similarity measurement between source support vector $\mathbf{v}_j^s$ and the labeled target data set $\mathcal{D}_l^t$, our optimal decision boundary can be obtained by solving the following optimization problem:

$$\begin{aligned} &\min_w \frac{1}{2}||\mathbf{w}||_2^2 + C\sum\nolimits_{i=1}^{|\mathcal{D}_l^t|}\epsilon_i + C\sum\nolimits_{j=1}^{M}\sigma(\mathbf{v}_j^s, \mathcal{D}_l^t)\overline{\epsilon}_j \\ &s.t.\ \ y_i(\mathbf{w}^T\phi(\mathbf{x}_i) - b) \geq 1 - \epsilon_i,\ \epsilon_i \geq 0, \quad \forall(\mathbf{x}_i, y_i) \in \mathcal{D}_l^t \\ &\qquad y_j^s(\mathbf{w}^T\phi(\mathbf{v}_j^s) - b) \geq 1 - \overline{\epsilon}_j,\ \overline{\epsilon}_j \geq 0, \quad \forall(\mathbf{v}_j^s, y_j^s) \in \mathcal{V}^s \end{aligned} \tag{4}$$

In CDSVM optimization, the old support vectors learned from $\mathcal{D}^s$ are adapted based on the new training data $\mathcal{D}_l^t$. The adapted support vectors are combined with the new training data to learn a new classifier.

For support vectors from the source data set $\mathcal{D}^s$, weight σ reduces the influence of those support vectors that are located far away from the new training samples in target data set $\mathcal{D}_l^t$.

Also similar to A-SVM [17], we want to preserve the discriminant property of the new decision boundary over the old source data $\mathcal{D}^s$, but our technique has a distinctive advantage: we do not enforce the regularization constraint that the new decision boundary is similar to the old one. Instead, based on the idea of localization, the discriminant property is only addressed over important source data samples that have similar distributions to the target data. Specifically, σ takes the form of a Gaussian function:

$$\sigma(\mathbf{v}_j^s, \mathcal{D}_l^t) = \frac{1}{|\mathcal{D}_l^t|}\sum\nolimits_{(\mathbf{x}_i, y_i) \in \mathcal{D}_l^t} \exp\left\{-\beta||\mathbf{v}_j^s - \mathbf{x}_i||_2^2\right\} \tag{5}$$

β controls the degrading speed of the importance of support vectors from $\mathcal{V}^s$. The larger the β, the less influence of support vectors in $\mathcal{V}^s$ that are far away from $\mathcal{D}_l^t$. When β is very large, a new decision boundary will be learned solely based on new training data from $\mathcal{D}_l^t$. Also, when β is very small, the support vectors from $\mathcal{V}^s$ and the target data set $\mathcal{D}_l^t$ are treated equally and the algorithm is equivalent to training an SVM classifier over $\mathcal{D}_l^t \cup \mathcal{V}^s$ together. This is virtually equivalent to the combined SVM described in section 3.1. With such control, the proposed method is general and flexible, capturing conventional methods as special cases. The control parameter, β, can be optimized in practice via systematic validation experiments.

3.4 BCRF Contextual Model

Another important branch of cross-domain learning method is the *prior* model. By applying the already trained models f^s from source domain $\mathcal{D}^s$ to the target domain $\mathcal{D}^t$, we can get a set of concept detection confidence scores for each target data $\mathbf{x}_i \in \mathcal{D}^l$. That is, each target sample $\mathbf{x}_i$ can be represented by a set of concept scores $\{f^s(\mathbf{x}_i)\}$. These concept scores form a concept feature space and based on which classifiers f^t can be trained using $\mathcal{D}_l^t$ for classifying $\mathcal{D}_u^t$. In this *prior* model, the source models f^s are used as prior knowledge to generate concept score feature vectors for learning new target classifiers.

In this work, we generalize our prior work on boosted conditional random fields (BCRF) [11] for cross-domain learning under this *prior* framework. BCRF aims to incorporate the inter-concept relationships (modeled by a conditional random field) to help detect individual concepts. The two-stage framework of BCRF makes it natural to generalize for cross-domain learning. In the first stage, detection scores of a large scale concept ontology (374 LSCOM) are generated from source model f^s learned with source data $\mathcal{D}^s$. Then in the second stage, these detection scores are used as new feature inputs, and through graph learning the target models f^t are learned using labeled target data $\mathcal{D}_l^t$ by considering the inter-conceptual relationships. Specifically, the joint conditional posterior probability of class labels is iteratively learned by the well-known real AdaBoost algorithm. SVM classifiers are used as elementary learners for each iteration.

The BCRF algorithm was a stand alone submission for the TRECVID2006 high-level feature extraction task [2]. In TRECVID2006, BCRF was used as a cross-concept learning method where both BCRF and baseline detectors were trained over TRECVID2005 development data set (referred to as source data in this paper). Also, in TRECVID2006 the BCRF algorithm was applied to 16 (out of 39) concepts automatically selected by a concept prediction criterion by taking into account both the strength of inter-conceptual relationships and the robustness of each baseline detector, i.e., a concept is predicted to be amenable to contextual fusion when its correlated concepts show strong detection power and its own detection accuracy is relatively weak [11]. In TRECVID2006, only 4 out of the 16 predicted concepts were evaluated by NIST, and 3 of these 4 concepts: *car*, *meeting* and *military-personnel*, show significant improvements of more than 20%, with the 4th concept showing no performance change (see Fig.(1) for details). The advantage of BCRF was further demonstrated by the evaluation over a separate validation data set, where prediction was very accurate – 13 out of the 16 predicted concepts showed significant performance gains while the remaining 3 did not show performance difference. Such significant gains and the high level of prediction accuracy are very encouraging, and confirm the effectiveness of context-based concept detection and the prediction method across data from different years. It also addresses the open issue concerning the inconsistent effects of contextual concept fusion found in many previous works.

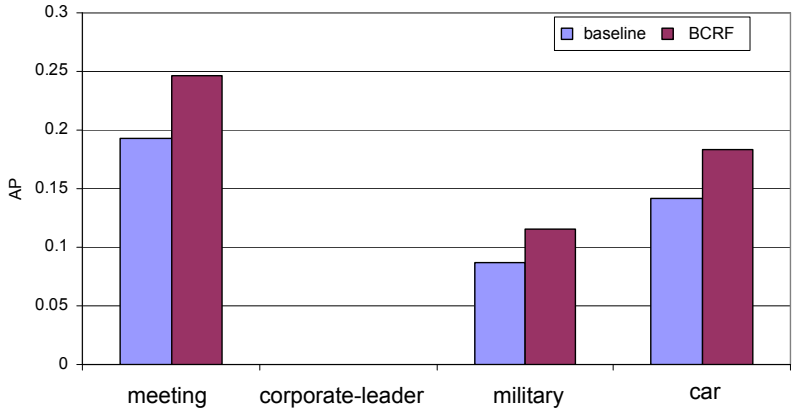

Figure 1: Performance of BCRF in TRECVID 2006 high-level feature extraction submission.

3.5 Multi-parameter Model Training

In training the baseline model over the new data set from TRECVID07, we experimented with a simple idea of fusing SVMs of multiple parameter sets, rather than choosing a single best parameter set. Grid search using n-fold cross validation has shown to be a reasonable choice for parameter selection under the assumption that the test data set is not too much different from the training data set. However from our empirical study, the single parameter selected by cross-validation usually is not the best parameter for detecting many concepts over test data. To alleviate this parameter selection problem, in TRECVID2007

we adopt a multi parameter setting model. Instead of training one model by a single parameter setting from cross-validation, we train multiple models with multiple parameter settings, and then fuse these models as final detectors. Specifically, in our work, the RBF kernel ($K(\mathbf{x}_i, \mathbf{x}_j) = \exp\{-\gamma||\mathbf{x}_i - \mathbf{x}_j||_2^2\}$) is used and there are two parameters to select for SVM construction, γ and C (Eqn.(1)). The procedure of the multi parameter setting model is as follows:

1. Empirically choose 10 different initial $\gamma = \{\frac{2^{-8}}{d}, \frac{2^{-6}}{d}, \frac{2^{-4}}{d}, \frac{2^{-2}}{d}, \frac{2^{0}}{d}, \frac{2^{2}}{d}, \frac{2^{4}}{d}, \frac{2^{6}}{d}, \frac{2^{8}}{d}\}$ and 4 different initial $C = \{2^4, 2^6, 2^8, 2^{10}\}$, where d is the dimension of the feature.

2. Through separate cross-validation comparisons, we have found ($\gamma = \frac{2^0}{d}$, $C = 2^4$) is a good initial choice.

3. Choose a set of parameter settings around the initial parameter setting. We chose 5 neighbors of C and γ and exhaustively pair these neighbors in the $\gamma - C$ space, i.e. 25 final parameter settings.

4. Using a the 25 parameter settings, train 25 SVM classifiers.

5. Fuse 25 SVM classifiers to generate the final ensemble classifier. Note the same parameter sets are used for every concept. Parameter search is no longer used.

The appropriate fusion strategy for multiple classifiers into an ensemble classifier is also an open issue, and in general it is not trivial to choose a fusion technique that works best for all concepts [6]. There are many possible fusion techniques, e.g. average, maximum, minimum, product, inverse entropy, inverse variance [6], and ensemble selection [7], and from the empirical study of many previous works the average fusion strategy usually generates robust performance. Thus for baseline models for the target domain, we simply utilize the average fusion method. To remove the influence of different scales of confidence scores from different classifiers, the classification scores of different SVMs are normalized first before average fusion by a sigmoid function: $\hat{f}(\mathbf{x}) = \frac{1}{1+\exp\{-f(\mathbf{x})\}}$.

3.6 Time and Model Complexity

Time and model complexity are also important factors to consider when choosing a cross-domain approach. Table 1 summarizes the data used for training and and an estimate for complexity and time usage.

Method	Train $\mathcal{D}^s$	Train $\mathcal{D}^t$	Complexity	Additional Training Time
apply source	all	none	$\lvert\mathcal{D}_l^t\rvert$	0x
retrain target	none	all	$\lvert\mathcal{D}_l^s\rvert$	1x
A-SVM	SVs	all	$\lvert\mathcal{V}^s\rvert + \lvert\mathcal{D}_l^t\rvert \approx \lvert\mathcal{D}_l^t\rvert$	1.25x
CDSVM	SVs	all	$\lvert\mathcal{V}^s\rvert + \lvert\mathcal{D}_l^t\rvert \approx \lvert\mathcal{D}_l^t\rvert$	1.25x
LSVM	regions	all	$\lvert\mathcal{D}_l^s\rvert * \lvert\mathcal{D}_u^t\rvert$	2x
standard combined	all	all	$\lvert\mathcal{D}_l^s\rvert + \lvert\mathcal{D}_l^t\rvert$	3x
replication	all	all	$3 * (\lvert\mathcal{D}_l^s\rvert + \lvert\mathcal{D}_l^t\rvert)$	9x

Table 1: Description of training data, complexity, and training time estimates for discussed approaches, assuming $|\mathcal{D}^s| > |\mathcal{D}^t|$. In TRECVID2007, $|\mathcal{D}^s| \approx 40k$ samples and $|\mathcal{D}^t| \approx 20k$.

Training on either source or target data alone is directly related to the amount of data in these domains (i.e. $|\mathcal{D}_l^s|$ and $|\mathcal{D}_l^t|$), defined here as $\mathcal{O}^s$ and $\mathcal{O}^t$, respectively. Similarly, a combined model uses both source and target data (i.e. $|\mathcal{D}_l^s| + |\mathcal{D}_l^t|$) in training so it's complexity is the combination of these complexities as well $\mathcal{O}^s + \mathcal{O}^t$. Other methods that seek to combine the different domains have different complexities depending on their approach.

Let $\mathcal{O}^{ts}$ represent the time complexity of training on combined source and target data. The LSVM approach needs to train $|\mathcal{D}_u^t|$ classifiers, one for each test sample. Thus the complexity of LSVM is about $|\mathcal{D}_u^t| * \mathcal{O}^{ts}$. In [4] the iterative training process for TSVM needs $P\mathcal{O}^{ts}$ complexity where P is the number

of iterations. Approximation methods can be used to speed up the learning process by sacrificing accuracy [4, 5], but how to balance speed and accuracy is also an open issue.

Replicated SVM training complexity is approximately a scalar of the combined approach. However, because it replicates features during training, its training scale to $2 * (N - 1)$ where N is the number of domains involved. In our experiment, only two domains were involved, but we are not aware of a limitation on the number of domains that could be included. One unique attribute about this particular model is that it hopes to have high performance across *all* included domains whereas the other approaches emphasize the target domain alone.

The CDSVM approach has relatively small time complexity. Let $\mathcal{O}^t$ denote the time complexity of training a new SVM based on labeled target $\mathcal{D}_l^t$. Since the number of support vectors from source domain, $|\mathcal{V}^s|$, is generally much smaller than the number of training samples in target domain, i.e., $|\mathcal{V}^s| \ll |\mathcal{D}_l^t|$, CDSVM trains an SVM classifier with $|\mathcal{V}^s| + |\mathcal{D}_l^t| \approx |\mathcal{D}_l^t|$ training samples, and this computational complexity is very close to $\mathcal{O}^t$.

As for BCRF, two SVM classifiers are trained during each iteration (see [11] for more details). So the time complexity of BCRF is about $2T\mathcal{O}^t$ where T is the number of iterations.

4 Analysis of TRECVID2007 Submissions

In this work, we evaluated several algorithms over different parts of the TRECVID data set [1]. The source data set, $\mathcal{D}^s$, is a 41847 keyframe subset derived from the development set of TRECVID2005, containing 61901 keyframes extracted from 108 hours of international broadcast news. The target data set, $\mathcal{D}^t$, is the TRECVID2007 data set containing 21532 keyframes extracted from 60 hours of news magazine, science news, documentaries, and educational programming videos. We further partition the target set into **training** and **validation** partitions with 17520 and 4012 keyframes respectively; in this partitioning process, we attempted to maintain equal coverage of broadcasts to avoid sample bias. The unlabeled target data, $\mathcal{D}_u^t$, is the entire TRECVID2007 test data set, for a total of 22084 keyframes from about 58 hours of broadcast video.

The TRECVID2007 data set is quite different from TRECVID2005 data set in program structure and production value, but they have similar semantic concepts of interest. All the keyframes are manually labeled for 36 semantic concepts, originally defined by LSCOM-lite [12], and in this work we train one-vs.-all classifiers.

For each keyframe, 3 types of standard low-level visual features are extracted: grid-color moment (225 dim), Gabor texture (48 dim) and edge direction histogram (73 dim). Such features, though relatively simple, have been shown effective in detecting scenes and large objects, and considered as part of standard features in high-level concept detection [1].

For all different algorithms, the RBF kernel, $K(\mathbf{x}_i, \mathbf{x}_j) = \exp\{\gamma||\mathbf{x}_i - \mathbf{x}_j||_2^2\})$, is used for all SVM classifiers. To avoid the difficulty of choosing one optimal parameter setting for the SVM classifier, a multi-parameter setting method is used. The basic idea is to train multiple SVM classifiers based on different parameter settings and then combine these multiple SVMs into an ensemble classifier. Through our empirical study, such a multi-parameter setting method usually provides robustly good performance. More details will be included in our detailed technique report.

4.1 Submission Definitions and Overall Performance

With six official submissions to compare, we chose to illustrate differences between several approaches that leveraged training on the source domain (TRECVID2005) versus those with training on the target domain (TRECVID2007). These official submission names and their purpose are defined in are defined in section 1 (repeated below). Fig. 2 illustrates the overall ranking and the average precision (AP) of our submissions and the order of the different submissions with respect to each other. Average precision is the precision

evaluated at every relevant point in a ranked list averaged over all points; it is used here as a standard means of comparison for the TRECVID data set.

- **A_COL_base6_T2_6** (R6): train on TRECVID2007 development data, multi-parameter models with average fusion across several modalities (referred to as *target baseline*)
- **A_COL_xd5_T5_5** (R5): train on TRECVID2007 development data and TRECVID2005 development data using feature replication, multi-parameter models with average fusion across several modalities (referred to as *xd* or *replication*)
- **A_COL_bcrf_base4_T7_4** (R4): trained with contextual scores from TRECVID2005 based models (using *columbia374*) and with TRECVID2007 development (referred to as *BCRF*) then average-fused with *target baseline*
- **A_COL_bcrf_xd_base3_T14_3** (R3): average fusion of *BCRF*, *target baseline*, and *xd replication*
- **A_COL_bcrf_xd_base_col3742_T16_2** (R2): average fusion of *BCRF*, *target baseline*, *xd replication*, and *columbia374*
- **A_COL_best_of_all1_T17_1** (R1): choose best-performing classifier for each concept over a validation data subset from all above submissions
- **col374**: not submitted, but publicly available model (see [16]) set covering 374 concepts in the LSCOM ontology and trained on only 60% of the TRECVID2005 development data (referred to as *source models*)

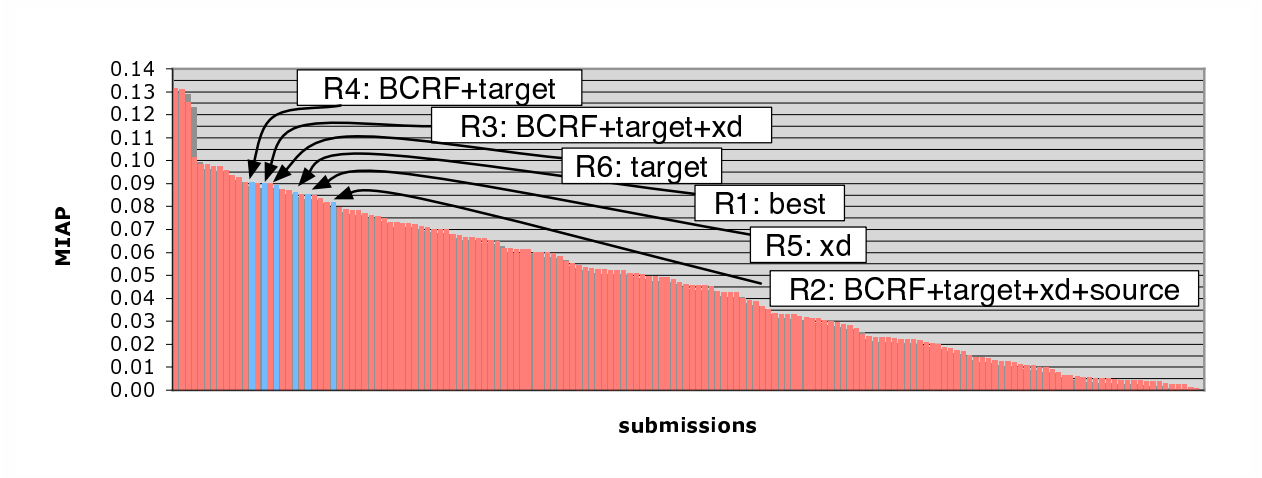

Figure 2: Standing among all submissions; our submissions are blue and all other group submissions are red.

We would like to note that although several methods are proposed in this paper, we only submitted a score set for the replication (or *xd*) approach. We provide a one-on-one analysis of the different proposed approaches in a supplemental empirical evaluation described in section 5.

Ordered by decreasing MIAP, or mean inferred average precision over all evaluated concepts, Fig. 2 also depicts a few trends among submission MIAP from which we can make empirical observations.

- R2 is the only run that directly fuses results from source model. The fact that it was ranked lowest among these runs indicates that indeed there is significant data difference between 2005 and 2007 domains, and thus blind application of source model is not a viable approach.

- When we compare the MIAP over 20 concepts, the specific cross-domain method (R5, feature replication) submitted is not as good as the target model (R6). However, as we will show in the next section, the cross-domain approach still shows noticeable gains for some specific concepts. Additionally, in a supplemental evaluation (section 5), other cross-domain approaches discussed in section 3.3 outperform the feature-replication cross-domain method we submitted in the official run.
- Similar to our findings in TRECVID2006, adding the BCRF model to utilize inter-concept context relations improves the overall performance (R4 is better than R6). Combination of context fusion (BCRF) and target models achieves the highest performance in our submitted runs.
- Selecting the best method by using a reserved data subset (R1) did not prove to be worthwhile based on the MIAP. This unreliable performance prediction could be due to the difference between the development and test data sets, and/or the limited size of the validation subset.

4.2 Specific Analysis by Concepts

Our TRECVID2007 submissions are briefly described in section 1 and results are shown in Fig. 2. A more specific analysis by concept is provided in Fig. 3. From these results, we can make a few important observations.

- The cross-domain (*xd*) approach (R5) provides benefits for some concepts that aren't available via target training (R6) in *maps, weather, and chart*.
- The cross-domain approach achieved performance comparable to retraining entirely new target models (*MIAP* or R5 and R6 differs by only 0.0045).
- BCRF, a contextual prior approach, provides complementary information during score fusion even though its individual performance may be weaker due to training on source model scores (R4 vs. R6).

While we did not create official submissions for all cross-domain approaches described in section 3.3.3, the observed performance indicates that a cross-domain approach to adapting models from prior data is both appropriate and necessary. We have conducted comparative studies of different cross-domain approaches using a reserved subset of development data of TRECVID2007. Details of such empirical studies will be described in the next section. Additionally, we found that contextual models are very useful even if they are constructed using models trained only on the source domain, as is the case of BCRF training on source domain concept scores and target domain labels, which is also known as a prior approach (see section 3.4). We also verified that a fusion of different approaches (i.e. BCRF and multi-parameter) increased average performance over the evaluated concepts. The next section of this paper detail additional experiments performed over the TRECVID2007 data set to better analyze the strengths and weaknesses of different cross-domain approaches.

5 Additional Empirical Cross-domain Analysis

In our supplemental empirical studies, we used the same source, $\mathcal{D}^s$, and target, $\mathcal{D}^t$ data set definitions presented in section 4, with one small exception: for the unlabeled target data set, $\mathcal{D}^t_u$, we use a subset of the official TRECVID2007 development data (the validation subset described in 4) instead of the official TRECVID2007 testing data. This choice was deliberate because the development data was fully labeled for all 36 evaluated semantic concepts, which avoids questions about full recall depth.

To guarantee model uniformity, for of the evaluated approaches, we train models with the same set of features (a concatenated 346-dim long feature vector to represent each keyframe) and a single set of parameter settings (an RBF kernel using $\gamma = \frac{1}{d}$ or 0.0029 for our experiments and $C = 1$, which are suggested

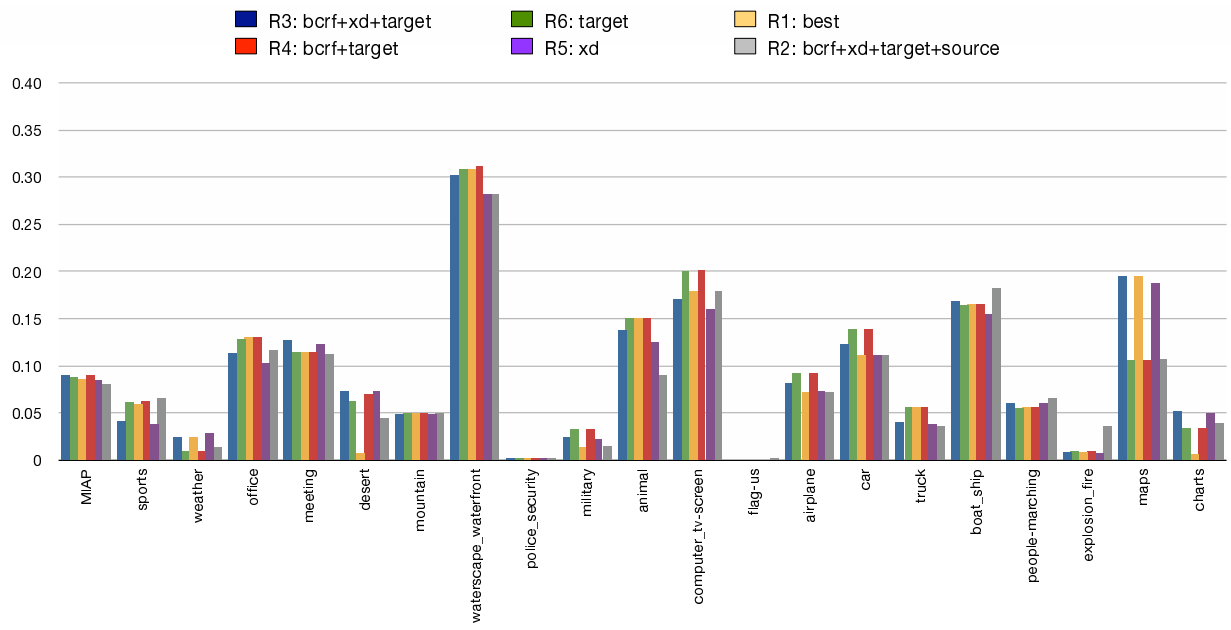

Figure 3: Inferred average precision of approaches versus concepts evaluated on the TRECVID2007 test set.

as default parameters in [3]). We used LIBSVM [3] for all SVM computations with a modification to include sample independent weights, as described in section 3.3.3.

5.1 Comparison of Methods

Comparing AP alone, the CDSVM method proposed in this work generally out-performs all other methods, as shown by Fig. 4. This is significant not only because of the higher performance, but also because of lower computation complexity compared to the standard combined, replication, and LSVM methods. Improvements over the target model and the combined model are particularly encouraging and confirm our assumption that a judicious usage of data from the source domain is critical for robust target domain models. Not all of the source samples are needed and inclusion of only source data support vectors is not overwhelming because each vector's influence is adequately customized.

5.2 Important Attributes of Approaches

While CDSVM has better average performance, further analysis demonstrates that it is not always the best choice for individual classes. Fig.4 gives the per-concept AP and is ordered such that frequency of positively labeled samples (as computed from $\mathcal{D}_l^t$) decreases from left to right. However, there are several trends seen in Fig. 4 that can be exploited to aide in the selection of the best approach on a per-concept basis. As hinted in the figure's concept ordering, categorizing the different concepts based on their positive frequency, D_l^t, provides a good preliminary grouping of best cross-domain choices. Positive frequency can be easily computed for either the source or target domain without any additional computation. We choose positive frequency because is directly related to the difficulty of a concept learning task, particularly in the case of discrete learning mechanisms, like SVMS. Another criterion available for a set of models trained only on source data, $|D^s|$, is the individual concept's performance relative to other concepts in a lexicon (here, the Columbia 374 [16]). While this metric is sensitive to the number of concepts in the lexicon, it is only

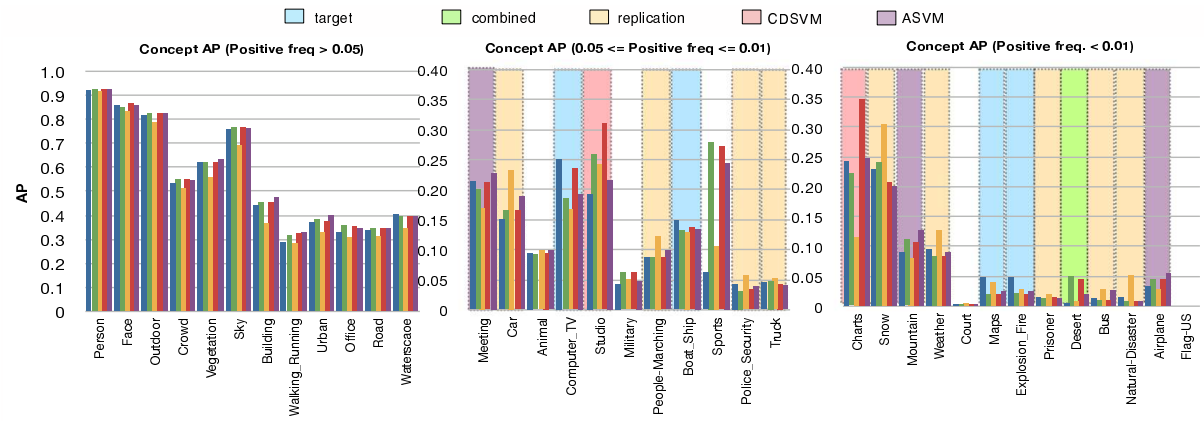

Figure 4: Average precision over a subset of $\mathcal{D}_l^t$ versus concept class for cross-domain methods; ordered by increasing frequency of positive $\mathcal{D}_l^t$ samples. Shaded concepts indicate the best method has a relative increase of at least 5% over all other methods.

needed as a coarse indicator for how well the model performed in the source domain. The use of these easily computable statistics is a subtle but important requirement that allows approach selection without any additional model learning or evaluation. In the next section, we describe a set of heuristic rules that can optimally select the best approach for each of the 36 evaluated concepts.

5.3 Predicting the Best Cross-domain Approach

Intuitively, CDSVM will perform well when we have enough positive training samples in both $\mathcal{D}^s$ and $\mathcal{D}^t$. It is highly probable that support vectors from $\mathcal{D}^s$ are complementary to $\mathcal{D}^t$, which can be combined with $\mathcal{D}^t$ to get a good classifier. However, when training samples from both $\mathcal{D}^s$ and $\mathcal{D}^t$ are few, positive samples from both source and target will distribute sparsely in the feature space and it is more likely that the source support vectors are far from the target data. Thus, not much information can be obtained from $\mathcal{D}^s$ and we should not use cross-domain learning. Alternatively, with only a few positive target training samples and a very reliable source classifier $f^s(\mathbf{x})$, the source data may provide important missing data for the target domain. In such cases, CDSVM will not perform well because target data is unreliable and instead the feature replication method, discussed in section 3.3.1 generally works well.

Based on the above analysis and empirical experimental results in Fig.4, a method predicting criterion is developed in Fig.5. With these prediction rules, we can increase our cross-domain learning mean AP from 0.263 to 0.271, but one must cautiously interpret these mean AP numbers. Though the overall mean AP improvement is relatively small (about 3%), the improvements over the rare concepts is actually very significant. If we compute the MAP over only concepts with lower frequencies, the improvement is as large as 22%.

6 Conclusions And Future Work

In this work we tackle the important cross-domain learning issue of adapting models trained and applied in different domains. We develop a novel and effective method for learning image classification models that work across domains even when the distributions are different and when training data is small. We also perform a systematic comparison of various cross-domain learning methods over the diverse and large-scale video data

if $(freq(\mathcal{D}_+^t) > T_1^t) \cup (freq(\mathcal{D}_+^s) > T^s)$ **then**
    Selected model = CDSVM
else if $AP(\mathcal{D}^s) > MAP(\mathcal{D}^s)$ **then**
    Selected model = Feature Replication
else if $(freq(\mathcal{D}_+^t) < T_2^t) \cap (freq(\mathcal{D}_+^s) < T^s)$ **then**
    Selected model = SVM over Target Labeled Set $\mathcal{D}_l^t$
else
    Selected model = CDSVM
end if

Figure 5: Method selection criterion. $freq(\mathcal{D}_+)$ is the frequency of positive samples in a data domain; $AP(\mathcal{D}^s)$ and and $MAP(\mathcal{D}^s)$ are the AP and MAP computed for source models on a validation set of source domain data; T_1^t, T_2^t and T^s are thresholds empirically determined via experiments.

set – the TRECVID2007 and TRECVID2005 data sets. By analyzing the advantages and disadvantages of different cross-domain learning algorithms, a simple but effective ruleset is proposed to determine when and which cross-domain learning methods should be used.

In terms of performance evaluation, on average, significant gains can be obtained by cross-domain learning algorithms over both the TRECVID2007 validation set and TRECVID2007 test set. This demonstrates the advantage of cross-domain learning for helping concept detection. As discussed in our empirical experiments, when our evaluation data set is very similar to the target training data set, CDSVM can significantly improve the detection performance by leveraging source information to help classify target data, with almost no additional time complexity. On the other hand, when the evaluation data is not so consistent with target training data, CDSVM may suffer from over-fitting and instead the feature replication approach that considers source and target data equally can learn a more balanced model.

Having confirmed the effectiveness of cross-domain learning methods, additional research can be done in three main directions:

- **Distribution similarity**: Motivated by the kernelized sample weighting employed in CDSVM (Eqn. 4) and the regularized constraints in A-SVM 3, we plan to explore additional ways to compute differences between domains. With a better way to determine differences, we could not only refine our heuristic set of prediction rules but also explore new approaches that reduce the magnitude of required new domain labels through user interaction or more a biased sample selection during training (similar to CDSVM).

- **Prediction refinement**: While the ruleset defined in 5 are adequate for this problem, we hope to include different metrics to refine this ruleset and eliminate heuristic thresholds. Additionally a richer macro-level, problem-driven prediction (as discussed above) can be added to the ruleset to adaptively predict which cross-domain learning algorithm to use for different evaluation data sets, based on the similarity of the data distributions from target domain and evaluation data set. Our experiments in section 5 demonstrate that even within the same domain, some approaches may be more fragile than others.

- **Unlabaled adaptation**: In this work we analyzed cross-domain approaches for labeled data only; the 36 concepts we considered had already been labeled in a massive group-driven effort by [13]. However, if we are to leverage the full strength of the LSCOM ontology [12], we must also look at approaches to adapt unlabeled concepts. While there has been some work in the speech community for this topic, there are many issues that are unique to the video and multimedia field.

7 Acknowledgements

We would like to thank Jun Yang and Alex Hauptmann for the availability of their software for the evaluation of the *ASVM* algorithm in our comparison experiments.

References

[1] S.F. Chang, *et al.*, "Columbia University TRECVID-2005 Video Search and High-Level Feature Extraction". In *NIST TRECVID workshop*, Gaithersburg, MD, 2005.

[2] S.F. Chang, *et al.*, "Columbia University TRECVID-2006 Video Search and High-Level Feature Extraction". In *NIST TRECVID workshop*, Gaithersburg, MD, 2006.

[3] C.C. Chang and C.J. Lin, "LIBSVM: a library for support vector machines", *http://www.csie.ntu.edu.tw/ cjlin/libsvm*, 2001.

[4] Y. Chen, *et al.*, "Learning with progressive transductive support vector machines", *IEEE Intl. Conf. on Data Mining*, 2002.

[5] H.B. Cheng, *et al.*, "Localized support vector machine and its efficient algorithm", *Proc. SIAM Intl' Conf. Data Mining*, 2007.

[6] Hsu, W., et al. "Discovery and Fusion of Salient Multi-modal Features towards News Story Segmentation". In *IS&T/SPIE Symposium on Electronic Imaging: Science and Technology - SPIE Storage and Retrieval of Image/Video Database*. 2004. San Jose, USA.

[7] Caruana, R., et al., "Ensemble selection from libraries of models" *Proceedings of the Twenty-First International Conference on Machine Learning 2004*. ACM Press: Banff, Alberta, Canada p. 18.

[8] H. Daumé III, "Frustratingly easy domain adaptation", *Proc. the 45th Annual Meeting of the Association of Computational Linguistics*, 2007.

[9] A. Gammerman, *et al.*, "Learning by transduction", *Conf. Uncertainty in Artificial Intelligence*, pp.148-156, 1998.

[10] W. Jiang, E. Zavesky, S.-F. Chang, A. Loui, "Cross-domain Learning Methods for High-level Visual Concept Classification", submitted *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Las Vegas, USA, April 2008.

[11] W. Jiang, S.-F. Chang, and A. C. Loui. "Context-based Concept Fusion with Boosted Conditional Random Fields". *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hawaii, USA, April 2007.

[12] M. R. Naphade, *et al.*, "A Light Scale Concept Ontology for Multimedia Understanding for TRECVID 2005," *IBM Research Technical Report*, 2005.

[13] Stphane Ayache and Georges Qunot, "Evaluation of active learning strategies for video indexing", *In Fifth International Workshop on Content-Based Multimedia Indexing (CBMI'07)*, Bordeaux, France, June 25-27, 2007.

[14] Smeaton, A. F., Over, P., and Kraaij, W. 2006. "Evaluation campaigns and TRECVid." *In Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*. Santa Barbara, California, USA, October 26 - 27, 2006.

[15] V.Vapnik and C. Cortes, "Support vector network", *Machine Learning*, vol.20, pp.273-297, 1995.

[16] Akira Yanagawa, Shih-Fu Chang, Lyndon Kennedy, Winston Hsu. "Columbia University's Baseline Detectors for 374 LSCOM Semantic Visual Concepts". Research Report Columbia University, March 2007.

[17] J. Yang, *et al.*, "Cross-domain video concept detection using adaptive svms", *ACM Multimedia*, 2007.

[18] Zhu Liu and David Gibbon and Behzad Shahraray, "Multimedia Content Acquisition and Procesing in the MIRACLE system," *CCNC 2006*, Las Vegas, Jan. 8-10, 2006.