Learning Heterogeneous Data for Hierarchical Web Video Classification *

Xian-Ming Liu
Harbin Institute of Technology
No. 92, West Dazhi Street
Harbin, P.R. China
xmliu@hit.edu.cn

Hongxun Yao
Harbin Institute of Technology
No. 92, West Dazhi Street
Harbin, P.R. China
h.yao@hit.edu.cn

Rongrong Ji
Harbin Institute of Technology
No. 92, West Dazhi Street
Harbin, P.R. China
rrji@hit.edu.cn

Pengfei Xu
Harbin Institute of Technology
No. 92, West Dazhi Street
Harbin, P.R. China
pfxu@hit.edu.cn

Xiaoshuai Sun
Harbin Institute of Technology
No. 92, West Dazhi Street
Harbin, P.R. China
xiaoshuaisun@hit.edu.cn

Qi Tian
University of Texas at San Antonio
One UTSA Circle
San Antonio, TX, USA
qitian@cs.utsa.edu

ABSTRACT

Web videos such as YouTube are hard to obtain sufficient precisely labeled training data and analyze due to the complex ontology. To deal with these problems, we present a hierarchical web video classification framework by learning heterogeneous web data, and construct a bottom-up semantic forest of video concepts by learning from meta-data. The main contributions are two-folds: firstly, analysis about middle-level concepts' distribution is taken based on data collected from web communities, and a concepts redistribution assumption is made to build effective transfer learning algorithm. Furthermore, an AdaBoost-Like transfer learning algorithm is proposed to transfer the knowledge learned from Flickr images to YouTube video domain and thus it facilitates video classification. Secondly, a group of hierarchical taxonomies named *Semantic Forest* are mined from YouTube and Flickr tags which reflect better user intention on the semantic level. A bottom-up semantic integration is also constructed with the help of semantic forest, in order to analyze video content hierarchically in a novel perspective. A group of experiments are performed on the dataset collected from Flickr and YouTube. Compared with state-of-the-arts, the proposed framework is more robust and tolerant to web noise.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

*Area chair: Nicu Sebe

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.

General Terms

Algorithms, Experimentation

Keywords

Transfer Learning, Web Video Classification, Hierarchical Taxonomy, Semantic Forest, Heterogeneous Data, Video Annotation, Video Search

1. INTRODUCTION

Coming with the explosive growth of Internet, video content analysis is a critical but challenging research topic, especially based on the difficulties to organize and search a huge scale of online videos. Different from an image, a video is difficult to understand because of its temporal relationship among each frame and the complex semantics. Especially, the video taggings are very expensive to fully acquired, resulting in the huge difficulty to train robust and effective video classifiers. In this work, we aim at investigating a more effective video classification / annotation strategy by learning video classifiers from the image domain.

Compared with the situations in classical video classifications, we have several new challenges when data is coming from the Internet. Regardless the low-level visual features and the semantic gap, the current video analysis methods usually suffer from several problems: Firstly, most video content analysis schemes require a certain amount of labeled data to train the models, and the labels are typically propagated in a top-down manner. For instance, each video shot is given a set of labels and the frames within this shot are inherent from these labels. This is usually unacceptable when dealing with huge scale of Web videos in terms of the labeling efficiency as well as the organization complexity. Secondly, instead of typically classifying the given tens or hundreds of categories, the Web videos are tagged with thousands or even more labels in total which are complicatedly correlative among each other. Therefore, traditional methods would not be capable to deal with this situation.

These drawbacks of the current approaches will cause i-

naccurate video contents understanding. In this paper, we are motivated at leveraging the cross-community knowledge across heterogeneous data towards automatic Web-based video classification. The cross-community data are used in two ways: firstly, a novel transfer learning algorithm is developed to make use heterogeneous auxiliary data; secondly, a hierarchical ontology, the *Semantic Forest* is proposed, which represents the semantics of the Web data and fits for transferred heterogeneously, from metadata of multiple communities in a data-driven manner. In more details, we collected large-scale video data from YouTube and Web images rom Flickr. Then by applying the proposed transfer learning algorithm, YouTube videos are classified using the knowledge learned from Flickr images. To integrate the classification results from frame-level into a higher semantic level, a bottom-up integration is performed based on a so-called *Semantic Forest*. Compared with the traditional approaches, our framework has the following novelties and contributions:

1. **Framework**: a novel Web video analysis framework is proposed in this paper by utilizing the cross-community heterogeneous data. By exploring the huge amount of user-labeled data from the Web, we are trying to figure out an efficient way of using these data, which is important for building video content analysis platform on a Web scale.

2. **Transfer Learning Strategy**: Though many transfer learning algorithms are proposed in recent years, there is still not an ultimate solution that works well in all scenarios. Therefore, it is still important to investigate the underlying principles for the task of classifying YouTube videos using Flickr images, and develop a specific strategy for this task, despite a lot of theoretical efforts have been made on this topic.

3. **Semantic Forest - Web Ontology**: Ontology plays an important role in video / image classification tasks when the category number gets large. We construct a hierarchical ontology named *Semantic Forest* in a data-driven approach and it organizes the semantic items within the Web data efficiently.

4. **Bottom-Up perspective**: To avoid the drawback in labeling video data, we view the video content analysis in a Bottom-Up perspective instead of the traditional Top-Down manner. The semantic of video clips is bottom-up integrated based on Semantic Forest, which has been quantitatively validated to be very effective later in this paper.

Above researches have covered several challenging research topics, including transfer learning, semantic representation, and ontology construction.

1.1 Related Work

Video classification and annotation is a hot research topic in this decade and a lot of methods are proposed in recent years [25][24][30]. TRECVID [1] provided an opportunity to the research on visual event detection and recognition on a standard dataset, and lots of approaches were built [2][6][7], most of which focused on finding better features or combinations and fused multiple classifiers together. Laptev [15] utilized the bag-of-word framework proposed by Sivic and Zisserman [22] and introduced a spatial-temporal local descriptor resulting good performance on a Hollywood Movie database, which is labeled using the movie scripts. However, all of these methods require large amount of labeled data, which is unrealistic in real application.

Some other work also aimed at dealing with the taxonomy and ontology problem, which is critical for classification and annotation tasks according to the observation that semantic items are correlated with each other. In the image domain, the ImageNet [9] provided an effective dataset and also the ontology for image classification and object recognition. While in video domain, few work focused on this topic. Yanagawa *et al.* [28] developed a set of baseline concepts detectors called Columbia374-baseline for 374 visual concepts chosen from LSCOM ontology and Yuen *et al.* also extended LabelMe to videos and constructed LabelMe Video [32]. A more recent work showed an alternative solution for automatic discovery and organization of descriptive concepts (labels) within large real-world corpora of user-uploaded multimedia in [3]. It used YouTube meta-data to build the classification taxonomy dynamically. Xie *et al.* also organized the concept forest in a probabilistic way, distinguished with traditional ways of treating semantic concepts as isolated nodes [27]. Besides, the concepts are correlated with each other, considering these correlations, Qi *et al.* proposed a CML method for video annotation, which obtained better performance than treating them independently as traditional approaches did [19].

As web communities such as Flickr and YouTube becoming popular, more researches concentrate on making use of these web sources to provide additional information to multimedia analysis. TRECVID has already included web videos in its current training and testing data since 2010. In [14], the author made use of Flickr images and tags calculating a new semantic metric for video classification. While in [23] and [26] YouTube metadata including user descriptions, comments and search logs were introduced to video classification and annotation procedure and provided more flexible and robust solutions for web videos. For TRECVID and consumer videos, web information was also incorporated to provide either additional training data or auxiliary information [10] [11]. Duan *et al.* [11] used transfer learning and pyramid matching to learn from YouTube video and then applied the classifiers to consumer videos.

To make better use the auxiliary web data, transfer learning [18] is naturally introduced. For both image and video analysis, lots of transfer learning approaches were built to adapt heterogeneous auxiliary data to the target domain, aimed at transferring knowledge learned from either documents or metadata to visual information. These approaches usually assume a bias in classification hyper-plane or assign a prior assumption from the source domain to the target domain. Yang *et al.* [29] proposed an Adaptive-SVM to transfer discriminative information from the source video domain to the target TRECVID domain. An Adaptive Multiple Kernel Learning (A-MKL) was also proposed in [11] to minimize both the structural risk functional and the mismatch between data distributions from the source and target domains, by extending Multiple Kernel Learning and Pyramid Matching methods. As discussed in previous part of this paper, the semantic concepts are correlated, the similar principle was introduced to transfer learning by Qi in [20] by transferring relational information across heterogeneous

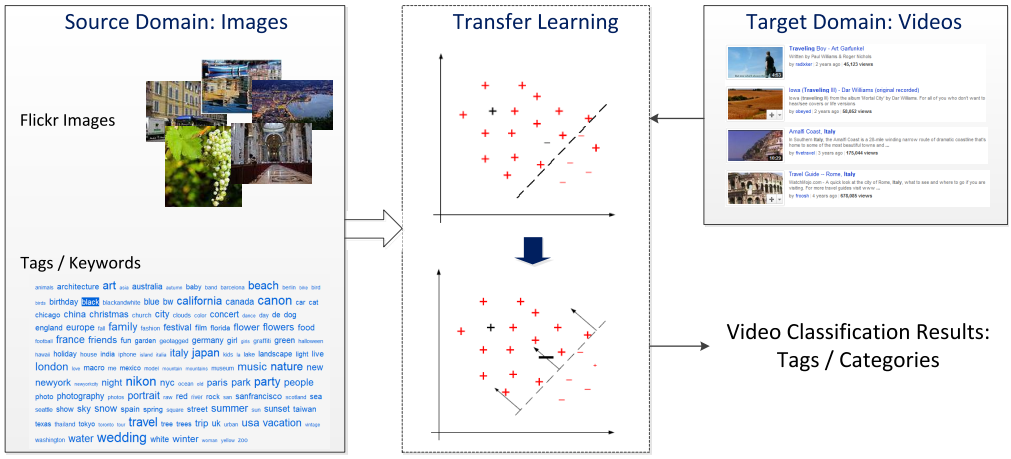

Figure 1: A candidate solution of web video classification and annotation by introducing additional sources, compared with traditional video classification framework.

data and applied in image classification. Similar issue was also discussed in [21]. More details about transfer learning please refer a survey by Yang *et. al* [18].

However, considering the following situation for video classification, the metadata or auxiliary data of web videos are utilized as additional information to help the classification tasks, labeled data is still hard to obtain. Using the auxiliary information such as tags and keywords as the labels of video clips will not only introduce extra noise, but also impossible to explicitly decide which shot or scene of the given video clip should have positive response to a specific tag.

A candidate solution is to use Flickr images as the source training data and transfer learning methods such as A-SVM [29] or A-MKL [11] are further adopted to transfer the discriminant information to the target YouTube video domain, as shown in Figure 1. A better result would be appreciated because the images are further incorporated as additional resource and hence provide more labeled training data.

Though positive, some problems still exist in this idea: Currently, the transfer learning strategies still fail to deal well with the heterogeneous situation, which is the exact one for transfer learning of taggings from from Flickr images to YouTube videos. Another crucial problem is that the existing works in heterogeneous transfer learning omits the concepts organization - the ontology or taxonomy.

1.2 Organization

Considering those problems mentioned above, a general framework of our solution is introduced in Section 2. By analyzing some statistical results from both Flickr and YouTube data, we discover the topic drifting phenomenon between heterogeneous data resources and introduce a novel transfer learning algorithm to annotate video frames by utilizing Flickr images' information in Section 3. Further, to cope with the complex semantic relation-ships, a hierarchical taxonomy named as Semantic Forest is mined from both video and image domains, based on which a bottom-up semantic integration is performed to build hierarchical semantic representation for Web videos in Section 4. In section 5, a group of experiments are designed on a large scale dataset to validate our framework and Section 6 concludes the whole paper.

2. GENERAL FRAMEWORK

A general framework of the proposed method is shown in Figure 2. It can be divided into two parts: 1) Transferring from Flickr images to YouTube videos; 2) Semantic Forest construction and bottom-up semantic integration for video clips. Compared with the prototype of candidate solution in Figure 1, two main modifications are shown in Figure 2. One is the novel transfer learning algorithm from the view of semantic concepts redistribution based on AdaBoost [12]. The other is the semantic integration procedure after the classification of video shots, which are the major contributions of this paper.

The system starts with extracting SIFT descriptors [16] within both the source Flickr images and YouTube videos. Then it constructs Bag-of-Words using K-Means clustering [22]. By analyzing the statistical distributions of semantic concepts over Bag-of-Words across from Flickr images to YouTube video key frames, a redistribution assumption is made and further we propose an AdaBoost-style transfer learning algorithm to transfer the discriminative information from Flickr images to YouTube videos, hence classify each video shot's key frames.

By adopting an unsupervised method over metadata from both Flickr and YouTube, a hierarchical ontology, the *Semantic Forest*, is built up according to tags' correlations. Finally, the semantic integration is performed to integrate the semantic concepts in low levels in a bottom-up manner, and thus describe the video clip content as a hierarchical semantic representation. To demonstrate our framework, an YouTube video classification application is set up between large-scale Flickr images and YouTube video clips. Some other potential applications also include hierarchical video browsing, web video tag suggestion, and keyword based video retrieval.

3. LEARNING FROM HETEROGENEOUS DATA

In this section, we want to answer the questions in the following three aspects:

1. Phenomenon: What's the difference of semantic concept-

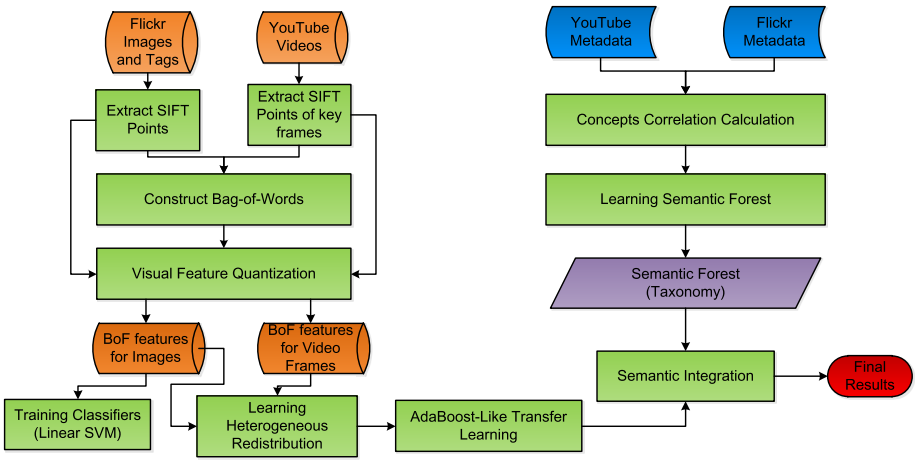

Figure 2: The Framework of the proposed method: Learning heterogeneous data for Web video classification

s distributions across different communities and media, explicitly from Flickr to YouTube?

2. Principles: Why the difference of distributions occurs and what's underlying this phenomenon? This could be principles for analyzing information transferring process between different communities.

3. Solutions: How to build efficient and effective algorithms to transfer discriminative information between heterogeneous data? Furthermore, how to connect different media sources and communities?

3.1 Heterogeneous Nature from Flickr to YouTube

In order to discover the underlying distribution principles of semantic concepts over different data resources, a group of statistical efforts have been done between collections of Flickr images and YouTube video clips. In this paper, we use the NUS-WIDE dataset [8], which contains 260,000 Flickr images and their corresponding tags as well as 2,000 YouTube videos collected from youtube by ourselves.

SIFT descriptors [16] on interesting points are extracted on both the entire Flickr image collection and the key frames of YouTube videos. Then K-means clustering is performed to construct Bag-of-Word following the classical work "Video Google" in [22]. Without loss of generality, we set the cluster number to be 500 as most approaches do, then adopt hard assign to assign each SIFT point to the nearest cluster center, which therefore achieves the vector quantification. Figure 3 shows this process. A major consideration in the BoW construction is whether or not to use the YouTube video data. YouTube videos are usually of lower quality than Flickr images and thus will add certain amount of noise. However, by incorporating them into the BoW construction will reduce the influence of domain shift. In this paper, we compare both strategies and the conclusion is to use both of them.

Each visual word in the dictionary is a compact representation of middle-level semantic concept according to [4][5] and also can reflect the semantic meaning [13]. We treat the visual words as the basic elements that semantic concepts are consisted of, and then analyze the distributions of semantic concepts on the visual words over both Flickr images and

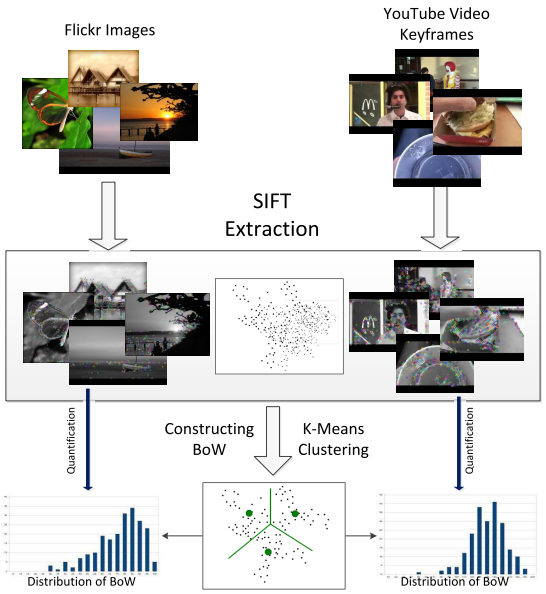

Figure 3: SIFT extraction and Bag-of-Word construction for Flickr images and YouTube video key frames.

YouTube video key frames. Let's denote Flickr image and YouTube video collections contain sets of semantic concepts as $T_{image} = \{t_{i,k}\}$, $T_{video} = \{t_{v,k}\}$ respectively. And further assume they share the same dictionary $\mathcal{D}$ (this is reasonable since they are spanned in the identical d-dimensional feature space $\mathcal{F}^d$). For each semantic concept t in both T_{image} and T_{video}, its distribution on the middle-level concepts $\mathcal{D}$ is defined as:

$$P(t|\mathcal{D}) = \sum_{s \in S_t} p(t, s|\mathcal{D}) \quad (1)$$

where S_t is the samples correspond to concept t. From an intuitive point of view, similar visual content should appear similarly even though within different data sources. However, our observation is totally different. To measure this similarity of two probability distributions of the same concept from Flickr and YouTube data collections respectively,

a balanced KL-Divergence is introduced as follow:

$$D_{KL}(t_{v,k}, t_{i,k}) = D_{KL}(t_{v,k} \| t_{i,k}) + D_{KL}(t_{i,k} \| t_{v,k}) \quad (2)$$

which means to what extent two probabilistic distributions are different. The detailed results are shown in Table 1. It is obvious that most concepts vary greatly in different domains, but they are still related with each other to some degree by distributed similarly.

So what are the reasons of this phenomenon? We consider following potential factors:

1. Descriptor: We use SIFT points in the feature detection and description phase, which is invariant to *scale*, *translation*, and *rotation*. Similar SIFT points from Flickr images and YouTube video frames are still able to be matched as shown in Figure 4. Thus, this factor is not one of the reasons.

2. BoW Construction: As illustrated in previous part, we use SIFT points from both Flickr and YouTube to construct the dictionary. And considering the data scale we utilized, the obtained Bag-of-Words representations are robust and stable. Thus, the BoW construction is not the reason either.

3. Heterogeneous Distribution: We compare some tags' distributions over the 500 visual words between Flickr images and YouTube videos, and some of the results are shown in Figure 5. It can be observed that there are significant probabilistic distribution shifts between different Flickr and YouTube for each concept. Therefore, we attribute the *Heterogeneous Phenomenon* to the redistribution of concepts.

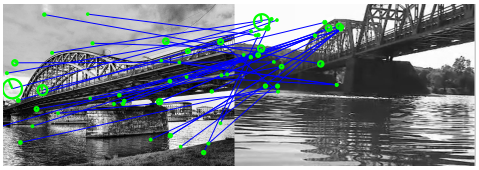

Figure 4: SIFT points matching between Flickr image and YouTube video frame. The left part is from Flickr, and the right one is from YouTube.

However, previous researches on transfer learning and multimedia retrieval usually omitted this fact and focused on transfer learning on feature space or decision hyper-plane. In this work, we focus on the re-distribution of mid-level concepts and make the following assumption:

LEMMA 1. (Redistribution of Semantic Concepts across communities)
Semantic concepts t_i and t_v from different communities sharing the identical feature space but different distribution over middle-level concepts ($p(t_v|\mathcal{D}) \neq p(t_i|\mathcal{D})$), is able to transfer to each other by redistributing middle level concepts as

$$p(t_v|\mathcal{D}) = \prod_{d_k \in \mathcal{D}} \eta_k^{v,i} p(t_i|d_k) = \eta \times P(t_i|\mathcal{D}) \quad (3)$$

where $\vec{\eta} = \{\eta_1, \eta_2, ..., \eta_{|\mathcal{D}|}\}$ is the redistribution parameter vector.

The lemma indicates that *the transfer learning can be viewed in a middle-level concepts' redistribution perspective*, which answers question 1 and 2 at the beginning of this section.

Table 1: Balanced KL-Divergence on the same tag between different media sources.

	Airport	Animal	Bear	Boats	Bridge	Buildings	Cars	Cityscape
Airport	0.1205							
Animal		0.0471						
Bear			0.0666					
Boats				0.1351				
Bridge					0.2068			
Buildings						0.0657		
Cars							0.1560	
Cityscape								0.2762

3.2 AdaBoost-Like Transfer Learning

According to the above analysis, a straightforward algorithm which transfers Flickr images' discriminative information to YouTube videos is shown in Algorithm 1 based on the *Redistribution Lemma*:

Algorithm 1:

Input: Linear classifier $f(x_i) = \mathbf{w_i} x_i$ for concept t
Output: Linear classifier $g(x_v) = \mathbf{w}_v x_v$ for concept t
foreach *visual word* $d_k \in \mathcal{D}$ **do**
  Estimate $p(t_i|d_k)$ and $p(t_v|d_k)$;
  Compute $\eta_k = \frac{p(t_v|d_k)}{p(t_i|d_k)}$;
end
$\vec{\eta} = \{\eta_k\}$;
$g(x_v) = \vec{\eta} \cdot f(x_i)$;
return $g(x_v)$

The basic idea of Algorithm 1 is to estimate the redistribution coefficients vector $\vec{\eta}$ from image domain to video domain, and hence re-organize the classifier $f(x_i) = \mathbf{w_i} x_i$ for concept t in image domain to adapt for the videos and obtain the target classifier $g(x_v)$. This is the direct conclusion from *Redistribution Lemma*. However, we want to impose some further considerations as follows:

1. Since the labeled video data is hard to obtain, it is impossible to fully and accurately estimate both $p(t_v|d_k)$ and the redistribution coefficient vector $\vec{\eta} = \{\eta_k\}$;

2. This algorithm assumes that all visual words are equally contributed to semantic concept t, which not holds all the time in the real case (where different concepts usually bias different visual words).

To solve these problems, we propose an AdaBoost-Like transfer learning algorithm as an extension of Algorithm 1. Equation (3) can be approximated by choosing the most representative visual words with the most discriminative information as follows:

$$p(t_v|\mathcal{D}) = \prod_{d_k \in \mathcal{D}} \eta_k^{v,i} p(t_i|d_k) \simeq \prod_{d_k \in \widetilde{\mathcal{D}}} \eta_k^{v,i} p(t_i|d_k) \quad (4)$$

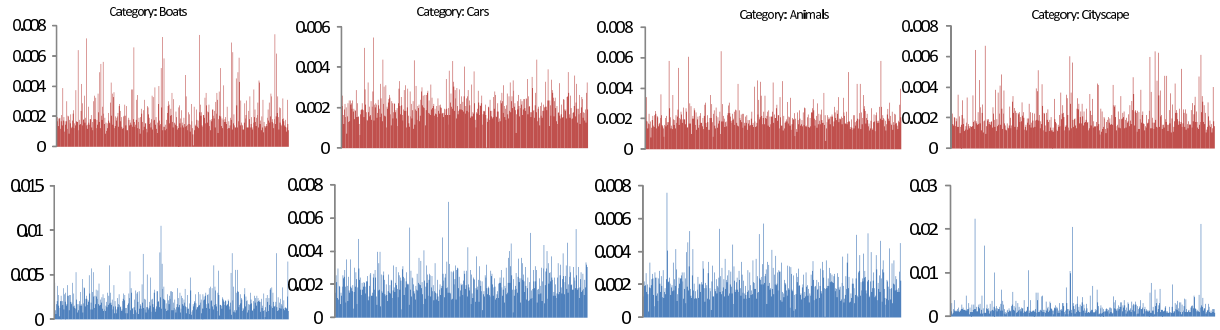

Figure 5: Semantic Concepts' distribution on visual words between Flickr images and YouTube video clips. The red ones are distribution of Flickr images on visual words, and the blue ones are the distribution of YouTube videos.

where $\widetilde{\mathcal{D}}$ is a subset of $\mathcal{D}$ with the most discriminative information. Then the transfer learning problems becomes a visual word selection and redistribution process, which can be easily solved by adopting the boosting idea. Thus, we naturally resort to the AdaBoost algorithm as shown in Algorithm 2.

Algorithm 2: AdaBoost-Like Transfer Learning Algorithm

Input: Linear Distribution $H_i = \mathbf{w_i} x_i$ for concept t, Loop L

Output: Linear Distribution H_v for concept t

Learn classifier $f(x_i)$ for images ;

Initialization: $H_v^{(0)} = H_i$, $\mathbf{w}_k = p(t_i|d_k)$, $W = \{w_k\}$;

for $l = 1, ..., L$ **do**

  Calculate training error on target domain according to $W^{(l)}$, $\epsilon^{(l)}$ on H_v^l ;

  Set $\beta_l = \epsilon^{(l)}/(1-\epsilon^{(l)})$;

  Update weight vector $w_k^{(l+1)} = w_k^{(l)} \beta_l^{1-\frac{1}{N}\sum |h^{(l)}(x)-y|}$;

  Update $w_k^{(l+1)} = \frac{w_k^{(l+1)}}{\sum w_k^{(l+1)}}$;

end

Output hypothesis $H_v = \sum w_K^{(l+1)} x_k$;

Apply hypothesis H_v on $f(x) \rightarrow f(H_v(x_v))$;

In Algorithm 2, the distribution $p(t_v)$ is initialized by the distribution on images collection $p(t_i)$, and the hypothesis for target domain $H_v^{(0)}$ is initialized with H_i from image domain. In each round of iteration, the middle level semantic concepts with lower response to target domain are weakened and the ones with higher response strengthened according to classifying error on the target domain samples. This is intuitive that the algorithm modifies the middle-level concepts' distribution to adapt the new domain.

In our application, we use linear-SVM trained on the Flickr image domain as the initial classifier, and this algorithm can be easily extended to any other type of classifiers. Besides, kernel methods are also available on this framework, by changing the hypothesis into kernel formulation. We use all the Flickr images to train the initial SVMs and in order to learn the hypothesis, a small part of YouTube videos are manually labeled and incorporated.

The AdaBoost-Like algorithm requires less labeled data in the target domain than Algorithm 1 because no probabilistic distribution on the target domain is needed to be estimated, and the training data is just used to evaluate the error of the current hypothesis. On the other hand, the AdaBoost-Like algorithm selects middle-level concepts by modifying the redistribution.

3.2.1 *Deep Insight of AdaBoost-Transfer*

There are mainly two concerns about the proposed AdaBoost-Transfer: the effect of emphasizing certain features and the robustness of AdaBoost methods. For the first concern, SIFT descriptors deal well with the scale change and are invariant to various situations. Thus, the detected SIFT points in both Flickr image domain and the YouTube video domain lie in the same feature space but distributed differently. The AdaBoost-Transfer aims at selecting the most stable feature under the unified hypothesis, and thus improves classification accuracy.

As to the second concern, the user tags on the Internet are noisy and not reliable enough, which may potentially lead to un-robustness of AdaBoost method. As addressed in the previous researches [31] and [20], this problem could be get rid of by using redundant data. In this paper, we introduce redundant data by using large-scale images and videos and thus eliminate the sensitiveness of AdaBoost algorithm.

3.2.2 *Relationship with Other Work*

The essence of this algorithm is to change the concepts' distribution over middle-level concepts followed by *Redistribution Lemma*, instead of assuming the distribution over raw feature space, which is the basic idea of A-SVM. Another related work is CCCL [20] which also use AdaBoost for transfer learning. However, CCCL chooses categories to involve relationships between different categories in each round. Our work is motivated by resolving the nature of domain transfer from Flickr to YouTube.

4. CONSTRUCTION SEMANTIC FOREST

4.1 Semantic Forest

Semantic Forest is a hierarchical cross-community taxonomy consists with a group of trees representing the ontology

for each category, which aims at building an efficient organization of semantic concepts covering various communities. To mine useful information from huge-scale metadata on the web, the construction of Semantic Forest follows a data-driven approach. This could benefit especially from the fact that the web is dynamic and changing all the time.

The principles in *Semantic Forest* construction are: The links between nodes are determined by their co-occurrences: More co-occurred tags tend to be connected more strongly in the taxonomy. The level of each node is determined by its importance, which is indicated by the frequency. We denote the tag collection and their correlations as $\mathbf{N} = \{n_i\}$ and $\mathbf{E} = \{e_{i,j} | n_i, n_j \in \mathbf{N}\}$, where

$$e_{i,j} = \frac{e_{i,j}^{Flickr} \cdot e_{i,j}^{YouTube}}{e_{i,j}^{Flickr} + e_{i,j}^{YouTube}} \tag{5}$$

is the harmonic-mean of the correlations between concept i and j in both Flickr and YouTube, and satisfying

$$\mathbf{e}_i^{Flickr} = \sum_{n_j \in \mathbf{N}} e_{i,j}^{Flickr} = 1, \mathbf{e}_i^{YouTube} = \sum_{n_j \in \mathbf{N}} e_{i,j}^{YouTube} = 1 \tag{6}$$

after normalization. This formulation considers both communities and response only when in both of the two communities the concepts are correlated with each other.

For each category of semantic concepts, we build a semantic tree representing the ontology within this category. After TF-IDF processing, useless tags are removed. Considering computation cost and data redundancy in the large scale web data, a direct but effective method is applied to extract semantic concepts approximately: Its basic idea is to first select the most important nodes in graph $G = (\mathbf{N}, \mathbf{E})$, then partition all nodes into sub-groups as their sub-trees, and finally select representative tags in each sub-tree. We iteratively perform this operation in each round until convergent. In each round, the nodes with top averaged degree is treated as more important and defined as:

$$\arg\max_i \frac{ei}{|e_i|_1}, e_i = \sum_j e_{i,j} \tag{7}$$

where the L1 norm is defined as the count of non-zero elements in each row vector $e_i = \{e_{i,j}\}$.

Considering a real situation that even single concept could make sense for multiple categories, for instance, "Apple" could be either a kind of fruit belonging to "Food" or a kind of computer (tag could be added to the semantic forest in multiple times). Algorithm 3 shows the detailed process of constructing Semantic Forest.

Following this process, the obtained Semantic Forest is a group of K-Trees. Other operations such as punch and trim can be easily integrated into the K-Trees.

Compared with former work mainly using linguistic models such as WordNet [17], the proposed *Semantic Forest* emphasizes on using a data-driven approach to automatically learn the representation and organization structure fitting for Web multimedia semantics, therefore is more appropriate for classifying Web videos and images than the traditional approaches. Another advantage is by using the large-scale data, the noisy tags are eliminated because of their small occurrences.

Algorithm 3: Constructing Semantic Forest

Output: Semantic Forest $\mathcal{F} = \{Tree_i\}$

Calculate $\mathbf{E}^{Flickr} = \{e_i^{Flickr}\}$ and $\mathbf{E}^{YouTube} = \{e_i^{YouTube}\}$;

Construct $\mathbf{E} = \frac{\mathbf{E}^{Flickr}\mathbf{E}^{YouTube}}{\mathbf{E}^{Flickr}+\mathbf{E}^{YouTube}}$ and $\mathcal{N}$;

$\forall Tree_i \in \mathcal{F}, Tree_i = Null$;

Select top K tags as roots for each tree following Eq. (8)

$r_i, r_2, ..., r_K$ And $\forall Tree_i \in \mathcal{F}, Tree_i = r_i$;

for $l = 1, ...L$ **do**

  foreach $Tree_i$ **do**

    foreach *node* t_i^l *on the* l*-th level of* $Tree_i$ **do**

      Find K best nodes related with t_i^l according to Eq. (5);

      Add these nodes as children of t_i^l;

    end

  end

end

4.1.1 Why Flickr and YouTube?

In this part, the reason why use Flickr + YouTube tags to construct the Semantic Forest is explained. Since the most popular tags on YouTube is related with music, movie, and pop singers, it is hard to extract effective information for video classification. Figure 6 shows an example of taxonomy tree extracted by only using YouTube information. Basically speaking, this is totally useless for efficient video annotation. However, the Flickr community does not focus on these aspects, and thus can remove useless information by using the harmonic average

$$e_{i,j} = \frac{e_{i,j}^{Flickr} \cdot e_{i,j}^{YouTube}}{e_{i,j}^{Flickr} + e_{i,j}^{YouTube}}. \tag{8}$$

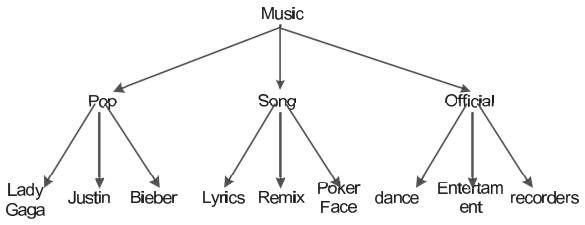

Figure 6: Taxonomy tree extracted only using YouTube information.

The most popular root concepts after filtering by Flickr become "Football", "Island" and "Africa" *et. al.* Due to the limitation of space, full semantic forest cannot be shown here.

4.2 Bottom-Up Semantic Integration

A bottom-up semantic integration is performed on the semantic forest, in order to integrate low-level semantic concepts into high-level ones. In our application of web video classification, this process is used to integrate the specific classification results of video key frames to more general concepts representation by using higher level semantics, including shots, scenes and video clips levels. All the operations are based on the assumptions that: classification

on each higher level concept is determined by not only itself, but also the children concept collection C. This can be formulated as:

$$\tilde{p}(A) = (1-\lambda)p(A) + \lambda \sum_{c_i \in C} p(c_i)p(A|c_i) \quad (9)$$

where $p(c_i)$ is the classification probability of child concept c_i, λ is a normalization factor to ensure that $\lambda \sum p(A|c_i) = 1$, $p(A)$ is the classification probability of concept A itself, and $\tilde{p}(A)$ is the final output result.

This integration is performed iteratively until reaching the top of each semantic tree. For each input video clip, the final classification result is the vector with every element representing classification score on corresponding semantic tree. This process is intuitively shown in Figure 7.

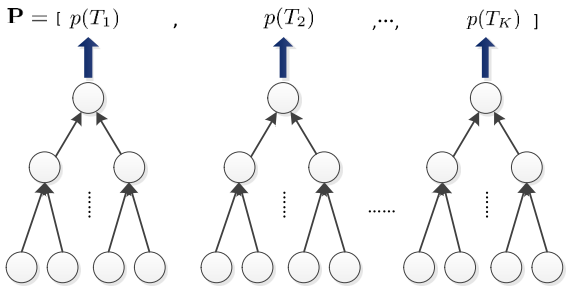

Figure 7: Classification result on semantic forest for the input video clip.

5. EXPERIMENTS

5.1 Preliminary

The data in our experiments include mainly two parts. For the image domain, NUS-WIDE dataset collected from Flickr is used, which contains 260,000 images, including corresponding tags. For the video domain, we collected 2,000 video clips from YouTube together with their keywords, descriptions and categories. 100 key frames are uniformly selected from each video clips without video shot detection, thus totally 200,000 video key frames are obtained. For the feature extraction, all Flickr images and 10,000 video key frames all over the dataset are chosen as input for K-Means clustering and construct the dictionary, by applying a Difference of Gaussian detector. Figure 8 shows examples of both images and videos.

To construct the Semantic Forest, we crawled 20,000 keywords using API provided by Flickr and Google YouTube respectively, and select most frequent 1,000 from these to construct the Semantic Forest. We set $K = 6$ in the construction which means each parent node has at most 6 children in the semantic tree. Thus the max depth of each semantic tree is 4.

5.2 Experimental Results

We compare various video classification scenarios on different setting and different algorithms in this section. 260,000 Flickr images and only 200 manually labeled YouTube video clips are used as training samples, by assigning each video with only one label. For the testing video examples, the keywords used to query on YouTube are stored as the ground truth. This leads to potential performance drop because not all the keywords are reasonable labeled by users.

Table 2: Performance Comparison with various methods

	Visual Word Construction	
	Flickr Only	Flickr & YouTube
VideoOnly	–	0.347
NoAdaBoost	0.283	0.314
AdaBoostTrans	0.304	0.326

Table 3: Experiments of the proposed method compared with other algorithms.

Category	VideoOnly	NoAdaBoost	AdaBoostTrans
Airport	0.212	0.054	0.072
Animal	0.048	0.138	0.136
Bear	0.417	0.417	0.306
Boats	0.072	0.344	0.469
Bridge	0.490	0.074	0.296
Buildings	0.911	0.465	0.854
Cars	0.373	0.630	0.451
cityscape	0.214	0.407	0.407

5.2.1 *Video Classification Performance*

Table 2 show the performance of the framework proposed in this paper compared with other methods, which are explained in details as follows:

VideoOnly - train a linear-SVM classier for each class using those 200 manually labeled videos as training samples and no Flickr images are incorporated in the training procedure.

NoAdaBoost - train a linear-SVM classier for each class using only Flickr images and no transfer learning strategy is used.

AdaBoostTrans - train a linear-SVM classier for each class using only Flickr images and labeled video data is used in the AdaBoost-Like transfer learning process.

The **VideoOnly** method uses only precisely labeled video clips as training data, which performs best but needs most human label. The performance of this method is the upper bound of video classification algorithms in our assumption.

Another observation is that the performance will improve by constructing dictionary using both Flickr images and YouTube video frames instead of using Flickr images only. This increases visual features' diversity and thus makes visual words more robust in representing visual content.

Table 4 shows detailed performance comparison on eight different categories. Since we don't incorporate motion feature into the training stage, all methods perform poorly on concepts such as "Animal" and "Airport". "Animal" videos are usually about moving animals beyond the ability of SIFT features, and "Airport" is really hard since airport is a very dynamic scene.

On the other hand, for some complex concepts such as "cityscape", we achieve better and more stable performance. This is because solely video data cannot provide sufficient information for these complex concepts, and large-scale Flickr images will make it much easier to solve this problem.

Figure 8: Examples of Flickr images and YouTube videos used in the experiments.

Table 4: Performance of semantic integration on Semantic Forest.

VideoOnly	0.347
AdaBoostTrans	0.326
AdaBoost-SemanticForest	**0.354**

An overall conclusion is that, the AdaBoost-Like transfer learning algorithm utilizes Flickr images to extend the representation ability of visual words, and also improves the stability of classifiers learned from image domain.

5.2.2 *Semantic Forest and Semantic Trees*

In this section, we emphasize on analyzing the effects of Semantic Forest and Semantic Trees.

The usefulness of taxonomy in classification, detection and recognition has been proved in many researches [17][23][19]. We apply semantic integration process on the extracted Semantic Forest, for which the classification results are shown in Table 4.

AdaBoost-SemanticForest - refers to adding learned semantic forest to AdaBoostTrans scenario to integrate semantic concepts in a bottom-up manner.

It is shown that the integration results are better than other approaches, even exceed the method directly using precisely labeled videos in training. This improvement can be attributed to the correlation and hierarchical structure of Semantic Forest, which integrates the classification results of related concepts to the target concept. Another question: Does the level of semantic trees affect the performance of video classification? In this subsection, we show some results on semantic integration using semantic trees with different levels. Figure 9 shows the average video classification precision of se-mantic integration using different semantic trees with different levels.

We change the size of children for each node (K) and the level of semantic forests L. The results are shown in Figure 9. More levels will incorporate more semantic concepts and thus improve performance, but will increase time cost as well. For the reason of balance, we set the level $L = 4$. The other issue we want to discuss is the number of children K. A larger K will integrate more concepts in fewer levels, thus when $K = 8$, the performance curve increase more sharply. But this will introduce non-relevant concepts into classification, based on which the performance cure slows down as the level of semantic forest grows.

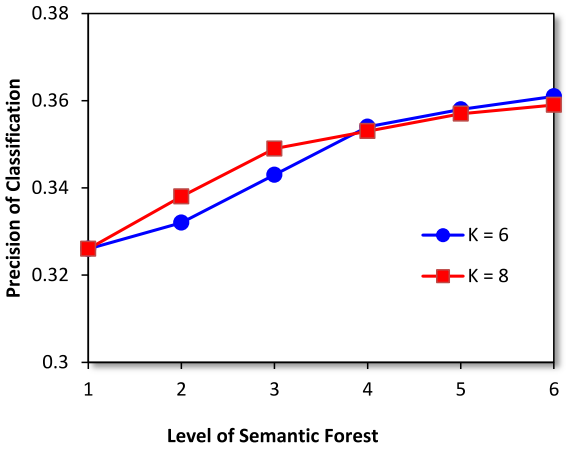

Figure 9: Performance of Semantic Forest integration under different depths and different children numbers.

6. CONCLUSION

This paper analyzes the heterogeneous phenomenon between different media. We assume the domain transfer from Flickr to YouTube is not caused by feature representation but the redistribution of middle-level semantic concepts. Subsequently, a simple but effective transfer learning algorithm is proposed using an AdaBoost-Like algorithm. Another contribution of this paper is the construction of the so-called semantic forest. To make use of keyword correlation into video classification, we build the semantic forest by mining concepts relationships within both Flickr and YouTube website. Based on this hierarchical taxonomy, a semantic integration operation is performed to integrate lower-level concepts to higher level ones. Several groups of experiments are performed on NUS-WIDE image database and 2,000 videos collected from YouTube.com. The experiment results show that our algorithm is both effective and efficient.

7. ACKNOWLEDGEMENT

This work was supported by National Natural Science Foundation of China (Grant No. 61071180 and Key Program Grant No.čž61133003), and in part to Dr. Qi Tian by NSF IIS 1052851, Faculty Research Awards by Google, FXPAL and NEC Laboratories of America, respectively.

8. REFERENCES

[1] Trecvid 2010 website. http://www-nlpir.nist.gov/projects/tv2010, 2010.

[2] A. Amir, J. Argillander, M. Campbell, A. Haubold, G. Iyengar, S. Ebadollahi, F. Kang, M. R. Naphade, A. P. Natsev, J. R. Smith, J. Tesic, and T. Volkmer. Ibm research trecvid-2005 video retrieval system. 2005.

[3] H. Aradhye, G. Toderici, and J. Yagnik. Video2text: Learning to annotate video content. In *IEEE International Conference on Data Mining*, pages 144–151, 2009.

[4] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *IEEE Computer Vision and Pattern Recognition*, pages 2559–2566, 2010.

[5] Y.-L. Boureau, J. Ponce, and Y. LeCun. A theoretical analysis of feature pooling in visual recognition. In *International Conference on Machine Learning*, pages 111–118, 2010.

[6] M. Campbell. Ibm research trecvid - 2006 video retrieval system. In *TREC Video Retrieval Evaluation*, 2006.

[7] S.-F. Chang, W. Hsu, L. Kennedy, L. Xie, A. Yanagawa, E. Zavesky, and D.-Q. Zhang. Columbia university trecvid-2005 video search and high-level feature extraction. In *TREC Video Retrieval Evaluation*, 2005.

[8] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Conference on Image and Video Retrieval*, pages 1–9, 2009.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li. Imagenet: A large-scale hierarchical image database. In *IEEE Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[10] L. Duan, I. W.-H. Tsang, D. Xu, and S. J. Maybank. Domain transfer svm for video concept detection. In *IEEE Computer Vision and Pattern Recognition*, pages 1375–1381, 2009.

[11] L. Duan, D. Xu, I. W.-H. Tsang, and J. Luo. Visual event recognition in videos by learning from web data. In *IEEE Computer Vision and Pattern Recognition*, pages 1959–1966, 2010.

[12] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. In *European Conference on Computational Learning Theory*, pages 23–37, 1995.

[13] R. Ji, H. Yao, X. Sun, B. Zhong, and W. Gao. Towards semantic embedding in visual vocabulary. In *Computer Vision and Pattern Recognition*, pages 918–925, 2010.

[14] Y.-G. Jiang, C.-W. Ngo, and S.-F. Chang. Semantic context transfer across heterogeneous sources for domain adaptive video search. In *ACM International Conference on Multimedia*, pages 155–164, 2009.

[15] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *IEEE Computer Vision and Pattern Recognition*, 2008.

[16] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004.

[17] M. Marszalek and C. Schmid. Semantic hierarchies for visual object recognition. In *IEEE Computer Vision and Pattern Recognition*, 2007.

[18] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22:1345–1359, 2010.

[19] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang. Correlative multi-label video annotation. In *ACM International Conference on Multimedia*, pages 17–26, 2007.

[20] G.-J. Qi, Y. Rui, Q. Tian, and T. Huang. Towards cross-cateogory knowledge propagation for learning visual concepts. In *IEEE Computer Vision and Pattern Recognition*, pages 897–904, 2011.

[21] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele. What helps where - and why? semantic relatedness for knowledge transfer. In *IEEE Computer Vision and Pattern Recognition*, pages 910–917, 2010.

[22] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *International Conference on Computer Vision*, pages 1470–1477, 2003.

[23] Y. Song, M. Zhao, J. Yagnik, and X. Wu. Taxonomic classification for web-based videos. In *IEEE Computer Vision and Pattern Recognition*, pages 871–878, 2010.

[24] M. Wang and X.-S. Hua. Active learning in multimedia annotation and retrieval: A survey. *ACM Transcation on Intelligent System and Technolgy*, 2, 2011.

[25] M. Wang, X.-S. Hua, R. Hong, J. Tang, G.-J. Qi, and Y. Song. Unified video annotation via multi-graph learning. *IEEE Transcation on Circuits and System for Video Technology*, 19, 2009.

[26] Z. Wang, M. Zhao, Y. Song, S. Kumar, and B. Li. Youtubecat: Learning to categorize wild web videos. In *IEEE Computer Vision and Pattern Recognition*, pages 879–886, 2010.

[27] L. Xie, R. Yan, J. Tesic, A. Natsev, and J. R. Smith. Probabilistic visual concept trees. In *ACM International Conference on Multimedia*, pages 867–870, 2010.

[28] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu. Columbia university's baseline detectors for 374 lscom semantic visual concepts. Technical report, Columbia University ADVENT, 2007.

[29] J. Yang, R. Yan, and A. G. Hauptmann. Cross-domain video concept detection using adaptive svms. In *ACM International Conference on Multimedia*, pages 188–197, 2007.

[30] Y. Yang, D. Xu, F. Nie, J. Luo, and Y. Zhuang. Ranking with local regression and global alignment for cross media retrieval. In *ACM Multimedia Conference*, pages 175–184, 2009.

[31] Y. Yao and G. Doretto. Boosting for transfer learning with multiple sources. In *IEEE Computer Vision and Pattern Recognition*, pages 1855–1862, 2010.

[32] J. Yuen, B. C. Russell, C. Liu, and A. Torralba. Labelme video: Building a video database with human annotations. In *International Conference on Computer Vision*, pages 1451–1458, 2009.