Taxonomic Classification for Web-based Videos

Yang Song, Ming Zhao, Jay Yagnik, and Xiaoyun Wu
Google Inc., Mountain View, CA 94043, USA
{yangsong,mingzhao,jyagnik,xiaoyunwu}@google.com

Abstract

Categorizing web-based videos is an important yet challenging task. The difficulties arise from large data diversity within a category, lack of labeled data, and degradation of video quality. This paper presents a large scale video taxonomic classification scheme (with more than 1000 categories) tackling these issues. Taxonomic structure of categories is deployed in classifier training. To compensate for the lack of labeled video data, a novel method is proposed to adapt the web-text documents trained classifiers to video domain so that the availability of a large corpus of labeled text documents can be leveraged. Video content based features are integrated with text-based features to gain power in the case of degradation of one type of features. Evaluation on videos from hundreds of categories shows that the proposed algorithms generate significant performance improvement over text classifiers or classifiers trained using only video content based features.

1. Introduction

With the astounding growth of videos on the Internet (such as YouTube [24]), organizing videos into categories is of paramount importance for improving user experience and website monetization. As much as its importance, web-based video classification poses serious challenges to computer vision researchers [25]. The difficulties are multifold, including large data diversity within a category [25], lack of manually labeled video data, and degradation of quality in some videos. The paper aims to tackle these issues.

A taxonomy is a tree of categories. Taxonomic classification is to categorize videos according to a pre-defined taxonomy. In contrast to the 11 YouTube categories used in [25], a taxonomy can include hundreds of categories. Figure 1 shows snippet of a taxonomy. Classification results from deeper nodes in the taxonomy can be propagated up to help classification of their ancestor nodes.

Taxonomic classification has attracted much attention in text documents classification [14, 10, 21, 2], and is emerging for image categorization [3]. However, there hasn't much work done on video classification. We postulate that one main reason is the lack of labeled training data. Compared to text-documents and images, video labeling requires more man hours, which makes it harder to obtain properly labeled video data, let alone for hundreds of categories in a taxonomy. On the other hand, there are plenty of labeled web-documents, e.g., through projects like Open Directory Project [14]. A good amount of images are collecting in ImageNet [3]. Then is it possible to take advantage of labeled text-documents and/or images to facilitate taxonomic video classification?

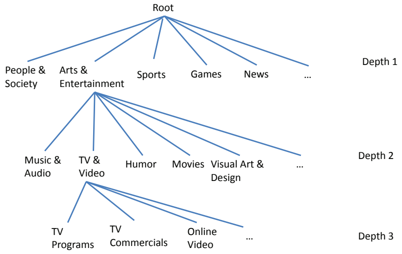


Figure 1. Snippet from a taxonomy.

Another challenge is that video quality may be degraded for Internet videos [24, 25]. However, those videos often have other information available, such as title, description, keywords, and associated search queries, which may not be available for consumer videos and surveillance videos. This inspires us to use different features and approaches for web video classification. How can we take advantage of the new information? How do we deploy the video-content-based audio/visual features in the mean time?

In an attempt to address these questions, a novel scheme is proposed to adapt classifiers trained on web-documents to video domain. For each video on YouTube, there are text-based information such as title, description, keywords, etc. However they are usually short, and have different characteristics from regular web-documents. We propose to use AdaBoost [6, 17] to adapt taxonomic classifiers trained from labeled web-documents to videos. This scheme also provides a natural way to deploy video-content-based au-

depth	1	2	3	4	5	total
num of nodes	27	274	583	140	12	1037

Table 1. Number of nodes (categories) at different depth levels in the taxonomy. The last column gives the total number of nodes.

dio/visual features in addition to the text-based classifiers. Various audio/visual features can be experimented.

To the best of our knowledge, this work is the first in adapting classifiers from web-documents to video domain, and is the first effort in classifying video into taxonomy of this scale (with hundreds of categories) using both text-based signal and video-content based features. Existing domain transfer work [4, 23] on videos are mainly transferring concepts (*e.g*. from "outdoor" videos to "office" videos) within the video domain.

The main contributions of this paper include the following. 1) A large scale video taxonomic classification system is presented, where the category taxonomic structure is utilized in training and evaluation; 2) To deal with the lack of labeled video data and take advantage of readily available labeled web-documents, a novel scheme is proposed to adapt web-documents trained taxonomic classifiers to video domain; 3) Video-content-based audio/visual features are integrated naturally with text-based features to gain power in the case of degradation of one type of features; 4) Various content-based features are experimented for effective video taxonomic classification.

1.1. Related work

The idea of using text with image features has been used in general video understanding [8, 20]. More recently, it is applied to video classification under multimodal framework [7, 16]. The existing works are limited to specific domains (e.g. [7] is on news video classification) or on a limited number of top level categories. Though sharing some spirits with [7], the way we adapt web-text classifiers to video domain is novel. It also provides a new scheme to combine text-based features with video content based features.

In [18, 25], video content based features are used to classify videos into top-level YouTube categories, and the reported classification results are not satisfying for practical deployment.

2. Overview

The taxonomy we use is similar to a subset of categories in Open Directory Project ([14]). Table 1 shows the statistics of the taxonomy. As shown in Figure 1, depth is defined as distance from the root.

Video taxonomic classification is to assign one or more categories in the taxonomy to a video. It is a multi-class and multi-label classification problem with hierarchical relationships between nodes (categories). In our approach, a classifier is trained independently for each node. Similarly, each classifier is applied independently to a video when doing classification. One reason for applying each classifier independently is that a video can have multiple labels, so it is not necessary to select one label against another. The hierarchical structure is deployed for generating training data, and for interpreting classification results.

For each node in the taxonomy, we have a pre-trained text-based classifier from labeled web documents. For categories without enough labeled video data (if there are any labeled video data at all), the original text-based classifier is used. For categories with enough labeled video data, we want to adapt the original pre-trained classifiers to video domain, and to integrate with video-content-based features. The next section explains these steps in detail.

3. Learning for video taxonomic classification

3.1. Use hierarchical structure of the taxonomy in training

The hierarchical structure of the taxonomy is used to generate training data for each category. In the document/video labeling process, a labeler is instructed to choose the deepest possible label in the taxonomy. For example, if */Arts & Entertainment/Movies* (a depth 2 category) and */Arts & Entertainment/Movies/Animated Films* (a depth 3 category) are both true for a video, the video is labeled as */Arts & Entertainment/Movies/Animated Films*. Multiple labels (not on the same path from the root) are allowed.

A "hierarchical one-against-all" strategy is used for training each category in the taxonomy. Let $\mathcal{C}$ denote the set of all the K categories in the taxonomy, $\mathcal{C} = \{C_1, C_2, \ldots, C_K\}$, where C_i is one category in the taxonomy. For a category $C \in \mathcal{C}$, let $\mathcal{A}(C) \subset \mathcal{C}$ be the set of ancestors, and $\mathcal{D}(C) \subset \mathcal{C}$ be the set of descendants. Category C is excluded from $\mathcal{A}(C)$ or $\mathcal{D}(C)$. Let $\mathcal{X}$ denote the set of N labeled samples, $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$, where x_j is one labeled sample. $L(x_j) \subset \mathcal{C}$ is the set of labels for sample x_j. Then, for a category C, the positive training set $Pos(C)$ and negative training set $Neg(C)$ are defined as:

$$\begin{aligned} &Pos(C) = \\ &\qquad \{x_j,\ s.t.\ L(x_j) \cap (C \cup D(C)) \neq \phi\} \qquad (1) \\ &Neg(C) = \\ &\qquad \{x_j,\ s.t.\ x_j \notin Pos(C)\ and\ L(x_j) \cap A(C) = \phi\} \end{aligned}$$

where ϕ is an empty set. Therefore, for each category, positive examples are all the videos labeled as either the category itself or one of its descendant categories. If a video has multiple labels, it is counted as a positive example for each category in its label set, and their ancestor categories. Negative examples consist of those videos which are not positive for the category itself and its ancestor categories. For each category, a classifier is trained using its positive and negative examples. This strategy applies to the training processes throughout this paper.

Taxonomy information is therefore embodied in classifiers through positive and negative training examples. This yields significant improvement over without using taxonomic structure.

3.2. Adaption from text documents to videos

The pre-trained text-based taxonomic classifiers include a classifier for each category in the taxonomy, trained by linear SVM (support vector machine) using labeled web-based text documents. In order to apply these classifiers to video domain, we propose the following domain adaption scheme. Each classifier in the pre-trained text-based classifiers is treated as a candidate "weak classifier" for video domain, and AdaBoost [6, 17] training is used to combine these "weak classifiers" into "strong classifiers" using labeled video data. In implementation, the K (K is the number of categories in the taxonomy) pre-trained classifiers are applied to each video so that a K-dimensional vector is obtained for each video, with one score from each pre-trained classifier. Values from this K-dimensional vector are then treated as candidate features/weak classifiers for AdaBoost. For each category with "enough" training data, a new AdaBoost classifier is learned. Decision stumps are used as weak classifiers. For a video x, the output of adapted classifier $H_C(x)$ for category C is given by,

$$H_C(X) = sign[\sum_{m \in \mathcal{M}_C} \alpha_m \cdot sign(\hat{h}_m(x) - t_m)] \quad (2)$$

where $\hat{h}_m(x)$ is the score from the pre-trained (linear SVM) classifier for category $m \in \mathcal{C}$. $\mathcal{M}_C$ is a subset of $\mathcal{C}$. $\mathcal{M}_C$, α_m, and t_m are selected/learned by AdaBoost using labeled videos for category C.

Text-based features for videos have different characteristics from those for web-based text documents. Therefore domain adaption is needed, which is verified in the experiments. One may also wonder, instead of domain adaption from pre-trained classifiers, can we use the text features extracted from videos directly to train a classifier? We postulate that domain adaption is superior. The main reason is that the number of labeled videos is often much less than the number of labeled web text documents. Applying the pre-trained text classifiers serves as one more layer of feature extraction. It is known [19] that effective feature selection can possibly significantly reduce model dimensions without impeding the performance of the learning algorithm, and even gain in generalization power by filtering irrelevant features. It is especially important when the number of training samples is small. These arguments are validated in experiments.

3.3. Integration of content-based features

We want to utilize video content-based features to improve classification performance. The domain adaption scheme described in section 3.2 provides a natural way to integrate content-based features and text-based classification results. The diagram in Figure 2 depicts the integration process.

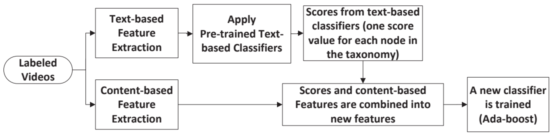

Labeled Videos
Text-based Feature Extraction
Apply Pre-trained Text-based Classifiers
Scores from text-based classifiers (one score value for each node in the taxonomy)
Content-based Feature Extraction
Scores and content-based Features are combined into new features
A new classifier is trained (Ada-boost)

Figure 2. Training diagram for adapting the pre-trained classifiers to video domain, and integrating with content-based features.

A content-based feature vector (as described in section 4.2) is extracted for each video, and is concatenated with the K-dimensional feature vector from the pre-trained classifiers (section 3.2). Values in this concatenated vector are used as candidate features/weak classifiers for AdaBoost training. The final classifier is similar to equation (2), with the candidates for $\hat{h}_m(x)$ extended to features from content-based feature extraction plus scores from the pre-trained classifiers.

Compared to concatenating content-based features directly with raw text features (instead of scores from pre-trained text-classifiers), the proposed method is advantageous. As argued in section 3.2 and validated in experiments (section 5.3), classification scores from pre-trained classifiers serve as more effective features than the raw text features. Moreover, the proposed method gives flexibility in using different types of classifiers for text-based features and content-based features. For example, linear SVM is used for the pre-trained text-based classifiers, and AdaBoost is used for content-based features and combining the content-based features with scores from pre-trained text classifiers.

4. Feature extraction

4.1. Text-based feature extraction

For each video, title, description and keywords are used to extract text-based features. Text features are represented by weighted text clusters, which are obtained from Noisy-Or Bayesian Networks [12].

4.2. Content-based feature extraction

Two types of global feature representations are used. The first type is to accumulate histograms across a video. The second is to use moments from time series multi-scale analysis.

4.2.1 Accumulated histogram across a video

“Histogram of local features” uses this scheme to generate global features for a video.

Histogram of local features. Laplacian-of-Gaussian (LoG) filters are exploited to detect interest points in each image frame (or at a certain down-sampling rate). For local descriptors, similar to SIFT [11], we compute 118 dimension Gabor wavelet texture features on the local region. These local descriptors are quantized according to a codebook. Histograms of codewords are accumulated across video frames to be used as features. The code-book is built by hierarchical k-means [13]. The size of the codebook is twenty thousand.

4.2.2 Moments from multi-scale analysis

The second type is to get feature values for each frame (or at a certain down-sampling rate), and corresponding feature values across frames form time series. Each feature corresponds to one time series. 1D Haar wavelet transform is applied to the time series at 8 scales, and moments (maximum, minimum, mean and variance) of the wavelet coefficients are used as features. The following content-based features use this scheme to generate global feature representation for a video.

Color histogram. A color histogram is computed using hue and saturation (4 by 4 bins) in HSV color space. Moreover, the mean and variance for each color channel are computed. Difference of the channel mean between center (a manually defined rectangle) and boundary areas of a video frame is also computed.

Edge features. Edges are detected by Canny edge detection [1]. The following edge features are computed: fraction of edge pixels, edge direction histogram, and the mean edge energy for pixels.

Histogram of textons. Texton histogram is computed as in [9]. The texton vocabulary is built by hierarchical k-means [13]. The vocabulary size is 1000.

Face features. Faces are detected by an extension of AdaBoost classifier [22]. The following face features are extracted: the number of faces and the ratio of largest face area to the image area.

Color motion feature and shot boundary feature. Color motion is measured by the Cosine distance of the color histograms of two consecutive frames. Shot boundary feature is a Boolean feature for each frame, indicating whether the frame is a shot boundary or not. A shot boundary detection algorithm similar to [26] is applied.

Audio Features. Audio features include the audio volume and coefficients of a 32-bin audio spectrogram [15] computed on a window around the video frame.

5. Experiments

Experiments are performed to verify arguments from previous sections, and to demonstrate the effectiveness of our algorithms.

5.1. Data

We use the taxonomy described in section 2. As described in 3.1, we use the deepest possible label along the path and allow multiple labels. Labels for 5789 videos are collected, resulting in 9087 labels. Table 2 shows how these labels distributed in different depths of the taxonomy.

Depth level	1	2	3	4	5
Number of labels	80	2079	4797	2057	74

Table 2. Number of labels for manually labeled videos at different depth levels in the taxonomy.

These labels covers 565 categories. However, as these videos are selected from YouTube by popularity(view count), the distribution across categories is not uniform. Figure 3 shows the number of videos belonging to each Depth 1 category, including videos from its descendant categories (equation (1)). 19 categories with more than 40 videos are shown. Category AE, “Arts & Entertainment”, has the most videos, followed by “Sports” and “People & Society”. See the caption of Figure 3 for category full names.

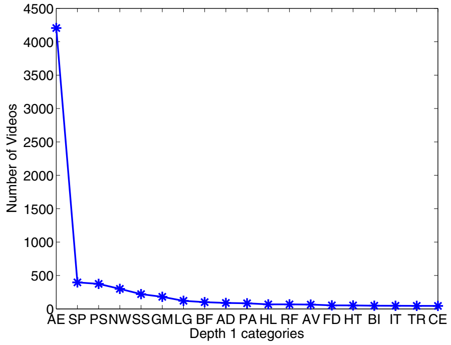
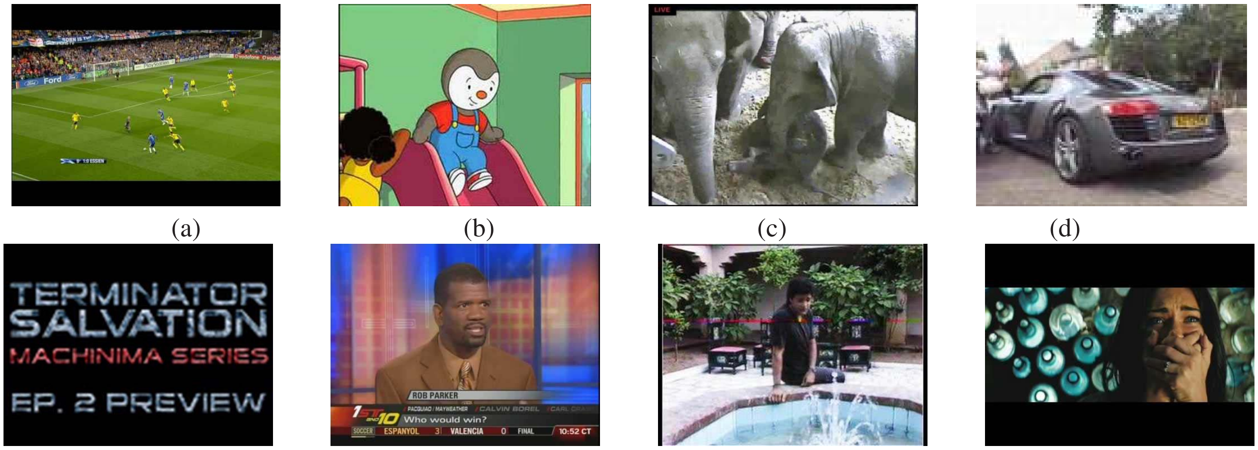


Figure 3. Number of videos in Depth 1 categories, using positive example definition as in equation (1). 19 categories with more than 40 videos are shown here. The full names of the categories are as following. AE: Arts & Entertainment; SP: Sports; PS: People & Society; NW: News; SS: Sensitive Subjects; GM: Games; LG: Law & Government; BF: Beauty & Fitness; AD: Adult; PA: Pets & Animals; HL: Hobbies & Leisure; RF: Reference; AV: Autos & Vehicles; FD: Food & Drink; HT: Health; BI: Business & Industrial; IT: Internet & Telecom; TR: Travel; CE: Computers & Electronics.

In the following experiments, these labeled video data are randomly split into two disjoint parts: 80 percent of



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 4. Sample videos from the evaluation set. Each video is represented by its thumbnail image on YouTube ([24]). Ground truth labels and classification results of these videos are given in Table 3. Links to these videos are in the format of http://www.youtube.com/watch?v=VIDEO_ID, where VIDEO_ID needs to be replaced by: (a) 0pSsFsKhrD0; (b) 1Xj2sG1loY0; (c) 3IPhEgbAp8U; (d) -dzo-1wMvqo; (e) 5W8kpP-Diks; (f) 6cIVtnzL0fw; (g) 6zTHwojAahk; (h) 98tPZCFL2TA. For example, link to video (a) is http://www.youtube.com/watch?v=0pSsFsKhrD0.

videos for training, and the other 20 percent for evaluation. Figure 4 shows sample videos from the evaluation set. Using positive example definition in equation (1), there are 80 categories with more than 40 videos as positive training examples.

5.2. Evaluation Measurements

As described in section 3, a classifier is learned for each category. In evaluation, these classifiers are applied to each video, and decisions are made for each category. Multiple categories (labels) can be assigned to a video. For a video x, let $\hat{L(x)}$ denote the set of computed labels from the classifiers, and $L(x)$ denote the set of ground truth labels. If C is a Depth-d category, a video x is true positive (TP) for category C if $C \in \{\text{Depth-}d\ ancestors\ of\ \hat{L(x)}\}$ and $C \in \{\text{Depth-}d\ ancestors\ of\ L(x)\}$; x is false positive (FP) for category C if $C \in \{\text{Depth-}d\ ancestors\ of\ \hat{L(x)}\}$ but $C \notin \{\text{Depth-}d\ ancestors\ of\ L(x)\}$. A Depth-$d$ ancestor of a category (label) is the Depth-d node (if it exists) in the path from the root to the category (including the category itself). True negative (TN) and false negative (FN) are defined accordingly. Precision (P) and recall (R) for a category are computed in the usual way, $P = TP/(TP + FP)$ and $R = TP/(TP + FN)$. F-score is the harmonic mean of precision and recall, $F = 2PR/(P + R)$.

5.3. Adaption to video domain and content-based feature integration

We have text-based classifiers trained via linear support vector machine ([5]) using 215,446 manually labeled web-based text documents, where the positive and negative example definitions as in section 3.1 are used in training. For brevity, we call these pre-trained text-based classifiers the “web-classifiers”. To achieve effective video classification, algorithms in sections 3.2 and 3.3 are applied to categories with more than 40 positive examples. 40 is a semi-arbitrary choice here.

Table 3 shows classification results on sample videos in Figure 4. For a video, if multiple labels are returned, only the deepest label along a path is displayed. For example, if both “/Arts & Entertainment” and “/Arts & Entertainment/Movies/Movie Reference/Movie Reviews & Previews” are returned from the classifiers, only “/Arts & Entertainment/Movies/Movie Reference/Movie Reviews & Previews” is shown in the table. This is consistent with our evaluation methods.

Figure 5 gives the performance by F-scores vs. different taxonomy depth levels. At each depth level, the average F-score for all the categories belonging to the level is computed. Figure 5(a) shows results of 80-category (with more than 40 positive examples) classifiers. Curves in Figure 5(b) are from 1037-category (all the categories in the taxonomy) classifiers. For categories without enough positive examples, web-classifiers are applied. There are no categories at Depth-5 with enough (more than 40) training examples, so the results in Figure 5(a) are only till Depth-4. The Depth-5 results from Figure 5(b) are from web-classifiers. In Figure 5, blue solid lines (with *) are from adapting web-classifiers to video domain and with content-based features; red dashed lines (with o) are from adapting web-classifiers to video domain but without content-based features; yellow

	Ground Truth Labels	Content-based Features only	Adaption Only	Adaption + Content-based Features
a)	/Sports/Team Sports/Soccer; /News/Sports News	/Sports; /News; /Arts & Entertainment;	/Sports/Team Sports/Soccer; /Arts & Entertainment	/Sports/Team Sports/Soccer; /News/Sports News
b)	/Arts & Entertainment/TV & Video/TV Programs/Children's Television; /Arts & Entertainment/Comics & Animation/Cartoons	/Arts & Entertainment/TV & Video; /Arts & Entertainment/Comics & Animation/Cartoons; /Games/Computer & Video Games	/Arts & Entertainment/TV & Video/TV Programs/Children's Television; /Arts & Entertainment/Comics & Animation; /People & Society/Kids & Teens/Children	/Arts & Entertainment/Comics & Animation/Anime & Manga; /Arts & Entertainment/Comics & Animation/Cartoons; /People & Society/Kids & Teens/Children
c)	/Pets & Animals/Wildlife; /Travel/Tourist Destinations/Zoos-Aquariums-Preserves	/Sensitive Subjects; /Hobbies & Leisure	none	/Hobbies & Leisure; /Pets & Animals
d)	/Autos & Vehicles/Vehicle Brands/Audi; /Autos & Vehicles/Custom & Performance Vehicles	/Arts & Entertainment; /Hobbies & Leisure	/Autos & Vehicles	/Autos & Vehicles
e)	/Games/Computer & Video Games/Shooter Games	/Arts & Entertainment/Movies/Movie Reference/Movie Reviews & Previews; /Arts & Entertainment/Offbeat; /Games/Computer & Video Games	/Arts & Entertainment/Movies; /Games/Computer & Video Games	/Arts & Entertainment/Movies/Movie Reference/Movie Reviews & Previews; /Games
f)	/News/Sports News; /Sports/Combat Sports/Boxing;	/Arts & Entertainment/TV & Video/TV Programs/TV Reality Shows; /News;	/News/Sports News; /Arts & Entertainment; /Sports/Combat Sports;	/News/Sports News; /Sports
g)	/People & Society/Family & Relationships/Romance; /Arts & Entertainment/Music & Audio;	/Arts & Entertainment/Music & Audio	/Arts & Entertainment	/Arts & Entertainment
h)	/Arts & Entertainment/Movies/Movie Reference/Movie Reviews & Previews; /Arts & Entertainment/Movies/Science Fiction & Fantasy Films	/Arts & Entertainment/Movies/Movie Reference/Movie Reviews & Previews;	/Arts & Entertainment/Movies/Movie Reference/Movie Reviews & Previews; /News; /Arts & Entertainment/Celebrities & Entertainment News;	/Arts & Entertainment/TV & Video/TV Commercials; /Arts & Entertainment/Movies/Movie Reference/Movie Reviews & Previews

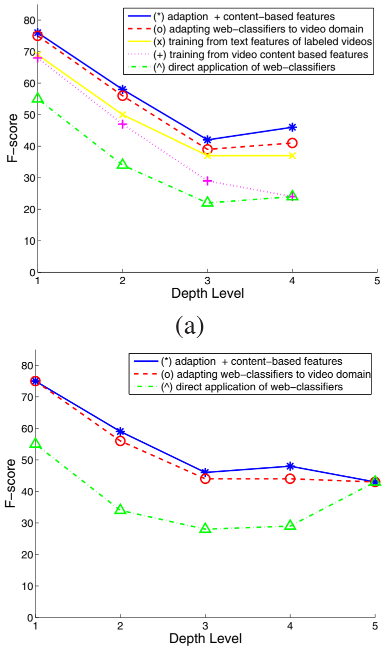
Table 3. Ground truth labels and classification results for sample videos in Figure 4. Each row corresponds to one video. The second column gives ground truth labels; the third column: results from classifiers trained using only content-based features; the fourth column: results from adapting web-classifiers to video domain; the fifth column: results from adapting web-classifiers to video domain and integrating with content-based features.

solid line (with x) is from classifiers trained using text features of labeled videos; magenta dotted line (with +) is from classifiers trained using video content-based features; green dash-dot lines (with triangles) are results of direct application of web-classifiers. From these curves, we can conclude that adapting web-classifiers to the video domain and with video content-based features works the best.

Figure 6 shows the classification performance for all the Depth 1 categories with at least 10 videos in evaluation. The categories with the best performance are "AE: Arts & Entertainment", "SP: Sports", and "PA: Pets & Animals".

5.4. Analysis of content based features

How useful are the video content based features? From Figure 5(a), F-score improvement of "adaption + content" over "adapting web-classifiers to video domain" is 1%, 2%, 3%, and 5% for depth 1, 2, 3, 4 categories, respectively. This improvement is desirable for a large scale system. Content-based features are important for videos with no text or noisy text or not-enough text information. For the video in Figure 4(a), "Adaption + content" gives better results than "Adaption only" (Table 3) because content features pick up the "/News/Sports News" category.



(a)

(b)

Figure 5. Performance comparison for different learning methods. **(a)**: 80-category classifiers; **(b)**: 1037-category classifiers. Blue solid line (with *): adapting web-classifiers to video domain and with content-based features; Red dashed line (with o): adapting web-classifiers to video domain but without content-based features; Yellow solid line (with x): classifiers are trained using text features of labeled videos; Magenta dotted line (with +): classifiers are trained using video content-based features; Green dash-dot line (with triangles): direct application of web-classifiers.

Various categories of content based features are described in section 4. Here we explore how much these features contribute to the classifiers. AdaBoost classifiers similar to equation (2) are used for adapting the web-classifiers to video domain and integrating with content-based features. One measurement of contribution of a feature type is the sum of stump-weights (α_m in equation (2)) of features selected from the feature type. Figure 7 shows the sum of stump-weights from classifiers vs. different types of features. Figure 7(a) is from "adaption + content-based features" classifiers; Figure 7(b) is from classifiers trained only using video content based features. The feature types are labeled along x-axis. "Web" is scores from the pre-trained web-classifiers, "SIFT" is "Histogram of local features", and the correspondence of the rest is directly from feature type names as in section 4.2.

Among visual features, "Histogram of local features", "Histogram of textons", and "Color histogram" contribute more to the classifiers than "Face features" and "Color motion feature and shot boundary feature".

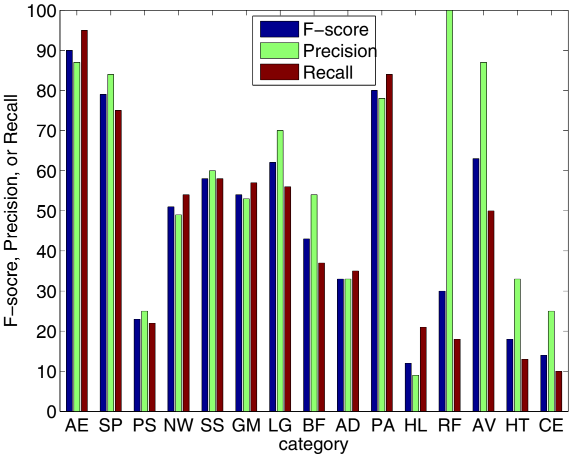
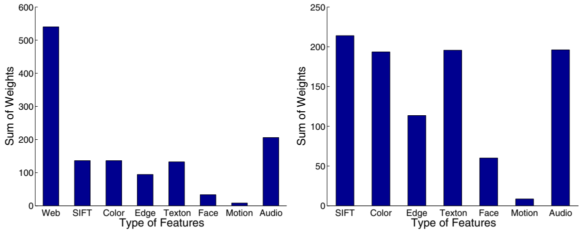


Figure 6. Performance for Depth 1 categories using 80-category classifiers. Results of 15 categories (with at least 10 evaluation videos in the category) are shown here. The performance is represented by F-score, precision, and recall. Category full names are the same as those in Figure 3. AE: Arts & Entertainment; SP: Sports; PS: People & Society; NW: News; SS: Sensitive Subjects; GM: Games; LG: Law & Government; BF: Beauty & Fitness; AD: Adult; PA: Pets & Animals; HL: Hobbies & Leisure; RF: Reference; AV: Autos & Vehicles; HT: Health; CE: Computers & Electronics.



(a)

(b)

Figure 7. Sum of stump-weights in the classifiers vs. different types of features. **(a)** is from "adaption + content-based features" classifiers; **(b)** is from classifiers trained only using video content based features. The feature types are labeled along x-axis. "Web": scores from the pre-trained web-classifiers; "SIFT": Histogram of local features; and correspondence of the rest is directly from the feature type names in 4.2.

6. Discussion and Conclusion

This paper presents a large scale video taxonomic classification system, which utilizes the category taxonomic structure in training and in interpreting the classification results. To take advantage of the available text information from videos on the Internet and to compensate for the lack of labeled video training data, a novel scheme is proposed to adapt the web-classifiers to video domain. Video content based features are integrated with text features to gain power in the case of degradation of one type of features.

Experiments show that the proposed algorithms generate significant performance improvement over the original text classifiers, and over the classifiers trained from using video-content based features. The performance has reached a satisfactory level for practical deployment.

We don't attempt to claim that we have solved the problem. However, our work sheds lights on the future directions. For example, another way to deal with the lack of training data is to use semi-supervised learning techniques [27]. Our approach is orthogonal to that. Combining our approach with semi-supervised learning methods can potentially gain more power in video classification.

Taxonomic structure is used to generate positive and negative examples. Taxonomy information is therefore embodied in classifiers via training data. This yields significant improvement over without using taxonomic structure. It may be worth exploring other uses of taxonomic structure (e.g. [2]) in future work.

Our work opens the path to use data from other domains to facilitate video classification. An algorithm for using labeled web-based text documents is presented in this paper. How to use data from image domain for video classification is an interesting topic to explore.

References

[1] J. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698, 1986. 4

[2] O. Dekel, J. Keshet, and Y. Singer. Large margin hierarchical classification. In *ICML*, 2004. 1, 8

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 1

[4] L. Duan, I. Tsang, D. Xu, and S. Maybank. Domain Transfer SVM for Video Concept Detection. In *CVPR*, 2009. 2

[5] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 2008. 5

[6] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, pages 119–139, 1997. 1, 3

[7] W. hao Lin and A. G. Hauptmann. News video classification using svm-based multimodal classifiers and combination strategies. In *ACM Multimedia, Juan-les-Pins, France*, 2002. 2

[8] A. G. Hauptmann and M. A. Smith. Text, speech and vision for video segmentation: The informedia project. In *AAAI Fall Symposium, Computational Models for Integrating Language and Vision*, pages 10–12, 1995. 2

[9] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision*, 43(1):29–44, 2001. 4

[10] T.-Y. Liu, Y. Yang, H. Wan, H.-J. Zeng, Z. Chen, and W.-Y. Ma. Support vector machines classification with a very large-scale taxonomy. In *SIGKDD Explorations*, 2005. 1

[11] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2004. 4

[12] R. E. Neapolitan. *Learning Bayesian Networks*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2003. 3

[13] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *CVPR*, 2006. 4

[14] OpenDirectoryProject. www.dmoz.org. 1, 2

[15] L. Rabiner and R. Schafer. *Digital processing of speech signals*. Prentice-Hall, Inc., 1978. 4

[16] C. Ramachandran, R. Malik, X. Jin, J. Gao, K. Nahrstedt, and J. Han. Videomule: a consensus learning approach to multi-label classification from noisy user-generated videos. In *ACM Multimedia*, 2009. 2

[17] R. E. Schapire. The boosting approach to machine learning: An overview. In *MSRI Workshop on Non-linear Estimation and Classification*, 2002. 1, 3

[18] G. Schindler, L. Zitnick, and M. Brown. Internet video category recognition. In *The First IEEE Workshop on Internet Vision, in CVPR*, pages 1–7, 2008. 2

[19] N. Slonim, G. Bejerano, S. Fine, and N. Tishby. Discriminative Feature Selection via Multiclass Variable Memory Markov Model. In *ICML*, 2002. 3

[20] M. A. Smith and T. Kanade. Video skimming and characterization through the combination of image and language understanding. In *CVPR*, 1997. 2

[21] A. Sun and E. Lim. Hierarchical text classification and evaluation. In *ICDM*, 2001. 1

[22] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *CVPR*, 2001. 4

[23] J. Yang, R. Yan, and A. Hauptmann. Cross-domain video concept detection using adaptive svms. In *ACM Multimedia*, 2007. 2

[24] YouTube. www.youtube.com. 1, 5

[25] S. Zanetti, L. Zelnik-Manor, and P. Perona. A walk through the web's video clips. In *The First IEEE Workshop on Internet Vision, in CVPR*, 2008. 1, 2

[26] H. Zhang, A. Kankanhalli, and S. W. Smoliar. Automatic partitioning of full-motion video. *Multimedia Systems*, 1(1):10–28, 1993. 4

[27] X. Zhu. Semi-supervised learning literature survey. Technical report, University of Wisconsin-Madison, 2008. 8