Multi-Modality Web Video Categorization

Linjun Yang[†], Jiemin Liu[‡], Xiaokang Yang[‡], Xian-Sheng Hua[†]
[†]Microsoft Research Asia
{linjuny, xshua}@microsoft.com
[‡]Shanghai Jiaotong University
{jemmy_2815, xkyang}@sjtu.edu.cn

ABSTRACT

This paper reports a first comprehensive study and large-scale test on web video (so-called *user generated video* or *micro video*) categorization. Observing that web videos are characterized by a much higher diversity of quality, subject, style, and genres compared with traditional video programs, we focus on studying the effectiveness of different modalities in dealing with such high variation. Specifically, we propose two novel modalities including a semantic modality and a surrounding text modality, as effective complements to most commonly used low-level features. The semantic modality includes three feature representations, i.e., concept histogram, visual word vector model and visual word Latent Semantic Analysis (LSA), while text modality includes the titles, descriptions and tags of web videos. We conduct a set of comprehensive experiments for evaluating the effectiveness of the proposed feature representations over different classifiers such as Support Vector Machine (SVM), Gaussian Mixture Model (GMM) and Manifold Ranking (MR). Our experiments on a large-scale dataset with 11k web videos (nearly 450 hours in total) demonstrate that (1) the proposed multimodal feature representation is effective for web video categorization, and (2) SVM outperforms GMM and MR on nearly all the modalities.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems—*video*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*Indexing methods*

General Terms

Algorithms, Experimentation.

Keywords

Video categorization, Web video.

MIR'07, September 28–29, 2007, Augsburg, Bavaria, Germany.

1. INTRODUCTION

With the rapid development of multimedia communication, sharing and authoring techniques, multimedia, especially video, services are becoming increasingly hot on the Web. A report on TechWeb [22] said that the revenue for web videos will "grow to $2.6 billion in 2009 from a projected $745 million this year." Many video sharing websites [1] [16] [26] [27] are getting popular and achieve more and more attentions and participation of people. From the report of Online Publishers Association [17], 5% of the Internet users watch web videos daily and 24% watch web videos at least once a week.

Facing such a large web video (here a web video means a *micro video* or *user generated video* from video sharing websites like YouTube) corpus, how to find a desired video or an interested category of videos are becoming increasingly difficult. This situation is similar to the one in searching for web pages. When the amount of web pages increases, people invented two ways to ease the process of information search: one is via text-based search engine; the other is by web directory [9], which can be generated automatically or edited by the web editors. These two approaches both fit for web videos. Currently, on most of the video sharing websites, category information is labeled by the user when he/she is uploading the videos. This manner has two disadvantages. First, the labeling will cost much human efforts and hence create a poor user experience. Second, the users have very different understandings of video categories, thus the labeled categories are often inconsistent. Automatic video categorization can solve these difficulties. It can ease the users from thinking and choosing the right categories for the uploaded videos, as well as the automatically annotated categories will be inherently more consistent. Besides, for those web videos which are not in the video sharing websites and whose category information is currently unavailable, automatic video categorization can also determine their categories.

There are some works on related topics such as *video genre categorization*, which is to categorize the videos into different genres such as "movie" and "news", and *video annotation*, which is to categorize video shots or video segments into different semantic concepts. Web video categorization is similar to these two topics in that they are all to categorize videos into some predefined concepts, categories, or genres. However, compared to them, web video categorization has two main differences. One is that web videos have high diversity in terms of subject, format, style, genre, quality and so on, compared with professional videos and home videos.

The other is that the category ontology is more flexible. Different from video genre categorization, web video category ontology concerns not only the "genre", but also the content, such as the category "animal" and "auto". Different from video annotation whose concept definition is at shot or segment level, video category is an attribute of an entire video. These two differences make web video categorization an equally, if not more, challenging problem.

Fortunately, web videos have rich information from multiple-channels. Not only visual and audio information, but also the surrounding text (the titles, descriptions and tags of web videos), even the social (i.e. the relationship among videos through the users or the recommendations) information can be applied. In the learning framework proposed in this paper, multiple modalities are incorporated, including visual low-level-feature modalities, which are widely used in existing video genre categorization; newly-proposed semantic modalities to handle the video content related categories; audio modalities; and surrounding text information in the web page affiliated with the web videos.

Given the information from multiple channels, how to represent them effectively remains a difficult problem. For the proposed semantic modality, two approaches are presented. One is to use the result of video annotation, i.e. the semantic concepts of every shot, and then a *concept histogram* is formed for each video. The other is based on *visual words* (similar to *visual terms* in [21]) which can be regarded as implicit concepts. Visual words are the cluster labels of shots achieved by clustering. The vector model of visual word is the semantic representation of the video. Further, LSA (Latent Semantic Analysis) [7] is employed to improve the vector model.

Text information in the web page affiliated with the web video is another important information channel for categorization. Among this information, titles, descriptions and tags are the most important and direct explanation for the corresponding video. In traditional ways of text retrieval, vector model is commonly employed as the feature representation. However, the web video texts have some intrinsic properties: (1) The surrounding text is extremely sparse. From our video corpus, average number of tags for one video is 4.35 and the dictionary of all tags contains thousands of terms. (2) To tag a specific category of videos, there are many candidate words. Such as the words "soccer," "sports" and "David Beckham" are normally used to tag a sports video. The video which is tagged "soccer" should be related to the video with tag "sports." In the sense of categorization, the videos having any one of such tags are similar. Motivated by such observation, a manifold ranking based propagated vector model is proposed to propagate the word-video similarity to other related terms.

In addition to these feature representations, we also study the performance of different machine learning algorithms which are commonly used in video annotation, including three paradigms: supervised discriminative learning, supervised generative learning and semi-supervised learning methods.

Based on the proposed multi-modality feature representations and the performance analysis on multiple classifiers, a comprehensive solution of web video categorization is proposed. In order to conduct a meaningful experiment on a large-scale dataset we collected videos from the web and labeled them with the category (the category definition will be described in Section 3.2 in detail). Finally a dataset consisting of 11333 videos labeled with 11 categories is formed. In such a relatively large dataset, the proposed categorization system combining multi-modality features and multiple classifiers is experimented.

In summary, there are three contributions in this paper: (1) This is the first study on web video categorization. The characteristics of web videos and web video categories are analyzed. The relationship between web video categorization and video genre classification, video annotation is investigated. (2) Multi-modality feature representation, especially video semantic features, are proposed. (3) A complete solution and system is presented and experimented on a large-scale web video dataset.

The rest of the paper is organized as follows. Section 2 will review related work, including video genre categorization and video annotation. We will study the characteristics of web videos and web video category definition in Section 3. In Section 4, various modality features are proposed. The system framework and employed machine learning algorithms for web video categorization are then introduced in Section 5. The performances of the proposed feature representations and the categorization system are presented in Section 6, followed by conclusion in Section 7.

2. RELATED WORK

As aforementioned, video genre categorization and semantic video annotation are two closely-related topics to our work.

2.1 Video Genre Categorization

To our knowledge, [8] may be the first work on automatic video genre categorization. They proposed a so-called three-step methodology to categorize films into news cast, sports (car race and tennis), animated cartoon and commercials. Firstly they used some low level features (including color statistics, motion vector, cut detection, audio statistics) as film syntactic. Then the syntactic are analyzed into some style attributes including object recognition, camera motion and etc. Finally a set of rule based classifiers and a classifier combination strategy are used to get the final result.

Most works in the following literatures use similar methodologies with the above, with differences in the selection of low level features, style attributes, or classifiers. Rasheed [19] proposed four new features, including average shot length, color variance, motion content and lighting key, from the viewpoint of film grammar. Their purpose is to categorize film previews into comedies, action, drama, or horror films. B.-T. Truong [25] designed a set of features from the study of editing effects, motion and colors used in different categories. The features include average shot length, percentage of each type of shot transitions, camera movement, and etc. Finally a decision tree is employed for classification.

From the above, it can be observed that most of the video genre categorization methods use only the video level statistical features, regardless of the video content. However, this will be insufficient for web video categorization, as the styles of the videos from a single category are frequently much more diverse. Hence, in our work, multi-modality analysis is employed.

Xun [28] revisited this problem from the viewpoint of ontology. A hierarchical video genre ontology is constructed and a hierarchical SVM classifier is designated to catego-

rize the video genres. Ontology is a promising methodology for such a task, but for web videos, it is more difficult to construct such an ontology definition because the categories are more diverse and a unified taxonomy is difficult to be defined.

2.2 Video Annotation

Video annotation is a main task of TRECVID [24], which is to categorize video shots into predefined semantic concepts, such as *person, car, sky, people-walking* and etc. We argue that video categorization, especially in the context of web videos, is similar to video annotation, except two differences: (1) video categorization has different category/concept ontology compared to video annotation, though some of the concepts could be applied on both; (2) video categorization is to categorize videos while video annotation is to categorize video shots or video segments. Besides, Video annotation and video categorization share similar methodology: Firstly low-level features are extracted, and then certain classifiers are trained and employed to mapping the features to the concept/category labels. Video annotation is regarded as a challenge problem [20]. Many of the methods, including supervised (such as SVM) and semi-supervised (such as Manifold Ranking) machine learning algorithms have been experimented on this problem [3] [5] [10] [11].

3. PROBLEM ANALYSIS

3.1 Characteristics of web videos

In order to categorize web videos, their characteristics should be firstly identified. Through our observation, web videos can be regarded as a hybrid of various videos, including:

- Professional videos with high quality. Here, professional videos mean films, music videos, or TV, that is, whatever videos that are created by professionals. Normally such videos are excerpted directly from long original videos or recorded from the TV, which often are of high visual quality.
- Professional videos with low quality. Sometimes, users photographed the TV screen to get the videos of one's favorites. Or the videos are re-encoded with high compressing ratio. Such videos are typically of low visual quality.
- Home videos recorded by camcorders. Such videos are the same as the conventional home videos.
- Home videos recorded by web camera. Such videos are of lower quality and narrower subject matter than the conventional home videos.
- Photo sequences with music. These are very common form for non-professional users to create one's own videos. Based on the observation, the amount of such videos is huge on the Web.
- Remix of video clips of the above types. Remix is a new and popular video service on the Web [12]. As a result, a lot of web videos are created by remix.

From the above summarization, it is obviously observed that web videos are of diverse characteristics: some videos are of high quality, others are not; some videos are created by professionals, others are not; some videos are of rich audio/music features (such as the professional videos and photo sequences with music), others are not. In addition, there are many videos that are remix of them. This makes web videos not only a hybrid collection, but also themselves the assembly of various forms. Moreover, a specific video category may contain videos of one or more types defined above, which makes web video categorization more challenging.

Besides the hybrid forms of web videos, there are other characteristics which are summarized as follows:

- The amount of web videos is huge and increases everyday. And the web video materials are easy to obtain.
- Web videos have rich social information. In the web page affiliated with the video, there are title, description, tags, comments, and recommendation information.
- Web videos are of varying qualities. Some videos are in high quality visually but the audio quality is poor. Some videos are of poor visual and audio quality while they are so popular and have rich user comments or tags.
- The duration and file size for web videos is limited by the administrators of most of the video sharing websites.
- The subjects of web videos are very rich.
- The video editing skills are quite different for different authors. The authors of web videos include professionals and non-professionals. Hence, unlike films, which follow the so-called film grammar, web videos, except the professional videos among them, are more ad-hoc without following film shooting and editing rules.

Web videos share some common features with home videos [14], but have some differences as well. As stated above, most of web videos are virtually home videos. In this sense, home videos are a subset of web videos. However, online home videos have some differences from traditional home videos: (1) Due to the quality loss incurred by compression, the quality of online home videos is lower; (2) The time-stamp data of home videos are lost when users upload them to the web; (3) The duration of the uploaded videos is limited; (4) Web videos have rich social information.

Based on the above analysis, it is natural to use multimodality analysis to facilitate web video categorization .

3.2 Video Category Analysis and Definition

Typically video categorization is only to categorize the genres of videos, such as film, cartoon, sports, news, and so on. For web videos, such genres are insufficient to differentiate their rich content. Category definitions excerpted from two popular video sharing web sites are listed in Figure 1 and 2.

From the above figures, it can be concluded that web video categorization is a different scenario from existing video categorization. Web video categories are more about video content than video genre, such as the category "animals" in Soapbox ("Pets & Animals" in YouTube). Even for some

Figure 1: Soapbox categories (http://soapbox.msn.com)

categories such as "music" which is also used in genre categorization, the definitions are different. In web video category definition, "music" means not only music videos but also the videos related to music, such as a concert or a girl is playing violin at home..

Currently, most of the video sharing websites support only one category for each video. We argue that this is insufficient since the categories are not mutually exclusive. One video can belong to "people" while it is also about "music," so it should be categorized into "people" and "music" simultaneously.

Based on the above observation and analysis we constructed a category definition which combines the categories from YouTube and Soapbox, discarding the categories which are difficult to define or very subjective, such as "Comedy." Then we employed 10 persons to label more than 10,000 videos collected from the web, with any number (0, 1 or multiple) of categories for a single video. After labeling, we again abandon some categories whose sample number is less than 100. The resultant categories are obtained, listed as below:

- *Auto & Vehicle*: the videos whose main content is about vehicles.
- *Commercial*: commercial videos.
- *Cartoon*: cartoon videos.
- *Animal*: the videos whose main content is about animals.
- *Entertainment*: the entertainment programs in the TV.
- *Sports*: the videos whose main content is about sports, not only sports video.
- *Film & TV*: films and TV plays.
- *People & Blog*: the videos which people use to record their life. Similar to home videos.
- *News*: news videos.
- *Music:* the videos whose main topic is about music, such as music videos, vocal concerts and even the non-professional music shows.
- *Game*: the videos via screen cast of computer games.

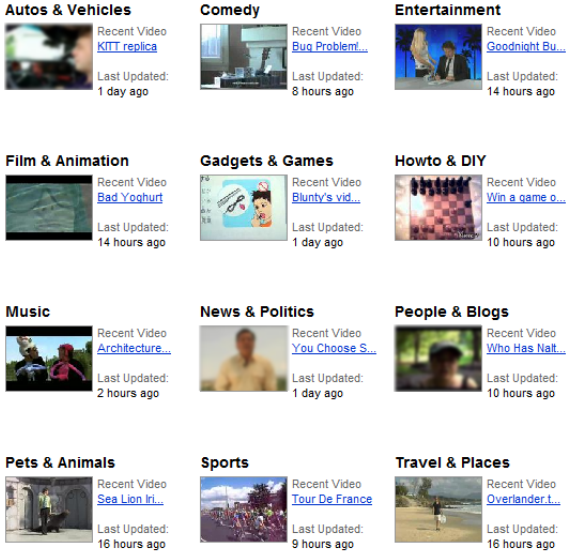

Figure 2: YouTube categories (http://youtube.com/categories)

Sample videos for each category can be found in YouTube and Soapbox, though some category definitions are not completely the same.

4. VIDEO REPRESENTATION

In order to categorize web videos affiliated with multiple information channels, we seek to find a multi-modality video representation. In Section 4.1 and 4.2 we will introduce the visual features, followed by audio features in 4.3. Thanking to Web 2.0, much social information is attached to videos by grassroots. In this paper, we only use surrounding text information including title, description and tags of videos. In the future, other information derived from the social networking consisting of both users and videos, such as video recommendation lists, can also be utilized for the purpose of categorization.

4.1 Visual low-level features

In order to investigate different characteristics of video categories at video level, we adopt the features in [28]: Temporal features including average shot length (1D), average color difference (1D) and camera motion (4D); Spatial features consisting of face frames ratio (1D), average brightness (1D) and average color entropy (1D).

Though such features can represent the entire video, they are insufficient in modeling the distribution of these video. Therefore, we add two more features: 2-order moment and 3-order moment besides the mean, for shot length, color difference, brightness and color entropy to better model their distribution. As a result, the video level features change from 9 dimensions to 17 dimensions.

4.2 Semantic features

As aforementioned, web video categories often reflect video content as well as the genre. Such as the "animals" and "autos" categories, they are defined from the viewpoint of video content. Intuitively the video semantic feature can be useful for such categorization. Here the video semantic feature

means the concept representation for the entire video. One simple representation of video semantic feature is to represent the video as the set of the concepts detected on every frame. Or it can be the concept detected on the key frame (or static thumbnail) for the whole video. Existing video annotation methods detect concept on every shot/subshot key frame, to represent the concept of the shot.

In this paper, two approaches to represent the video semantic feature are introduced. One is based on video annotation, and the other is based on visual words. Visual words are formed from video frame clustering, which represent implicit semantics. Both of them detect concept or visual word at shot level. Then a histogram or vector model is employed to represent the video semantic based on the set of concepts or visual words of the shots in the video.

4.2.1 Concept Histogram

TRECVID [24] includes a task named high-level feature extraction, also named as concept detection or video annotation. TRECVID 2006 defined a collection of 39 concepts and conducted the annotations on about 80 hour dataset. To alleviate the human efforts for concept labeling, we adopt this dataset to train the concept models. However, not all of the 39 concepts are appropriate to describe the content of web videos, such as "prisoner". Hence, we chose sixteen concepts from them, which we believe are appropriate for web video categorization, including *Building, Car, Face, Entertainment, Meeting, Military, Mountain, Office, Person, Vegetation, Road, Sports, Government-Leader, Sky, Waterscape_Waterfront and Studio* [24]. A more sophisticated concept list for web video categorization can be constructed by analyzing the distribution of semantic concepts in these videos, which is also our future work.

The concepts for every key-frame of a shot are predicted using SVM [4] with pre-trained concept models from the TRECVID dataset. After concepts are all detected for all the shots in the video, video concept histogram is formed, which is a 16-bin vector and every dimension of which denotes the occurring frequency of the corresponding concept in the video.

4.2.2 Visual Word Vector Model

To achieve a sound performance from the above concept histogram feature, specifically designed concept ontology for web videos and corresponding annotations should be made. Though we have chosen 16 concepts from TRECVID 39 concepts, they are far from our requirements in two senses: (1) such concepts are insufficient to describe the high-diversity web videos; (2) the concept ontology defined in TRECVID is not the best for web videos since it is specifically designed for professional TV programs.

Alternatively, in this section, we will propose a method that can automatically construct concept "vocabulary" from data, which is called *Visual Word* (similar to "*Visual Term*" in [21]) in the sense that the implicit words can be constructed from the visual features. The approach can be briefly stated as follows: (1) clustering. The key frames of all the videos are clustered into N clusters using any existing clustering algorithm. Then every key frame can be assigned a cluster label according to the clustering results. (2) For each video, the cluster labels of all its key frames can form one representation for it.

Algorithm 1 Construction of Visual Word Vector Model

1: Parse the video, get the shot key frames. Extract the features for all key frames. (In our experiment, block-wise LAB color moment is used.)
2: Cluster the key frames of all the videos in the dataset into N clusters. Any existing clustering algorithms can be used. In the experiment, K-Means is employed.
3: For every video, the visual word vector model representation is $[v_1, v_2, \ldots, v_N]$ with $v_i = \frac{\sum_j K_j=i}{M}$ where K_j is the cluster label for the j^th key frame and M is the total number of key frames for a video.

After the visual word representation is constructed, an intuitive idea is to treat the visual words as terms and the video as a text document. Therefore, the vector model (which will be described in detail in section 4.4) in information retrieval can be naturally adopted to solve this problem. The complete algorithm is summarized in Algorithm 1.

4.2.3 Visual Word LSA

Latent semantic analysis (LSA) [7] is a classic approach used in text information retrieval. F. Souvannavong et al. [21] proposed to adopt LSA in video shot retrieval system based on image region features and achieved good performance.

In our work, similar to [21], LSA is employed as a post-processing that is supposed to discover the semantic relationship between the visual words and therefore improve the performance of visual word vector model.

Given the video semantic feature represented as vector model $\overrightarrow{v} = [v_1, v_2, \ldots, v_n]'$, where v_i denotes the frequency of visual word k_i occurring in the video document. For all the m videos, a $n \times m$ term-document co-occurrence matrix can be constructed. Singular value decomposition of D gives $D = UEV'$ where $U'U = I$ and $V'V = I$. The relationship between video documents and the one between visual words can be discovered in the matrix U and V.

Through matrix U, the visual word vector model of the query video $\overrightarrow{q}$ can be transformed into a latent semantic representation $\overrightarrow{p} = EU'\overrightarrow{q}$. The dimensionality of $\overrightarrow{p}$ can be reduced by choosing only the d column vectors of U corresponding to the d largest singular values. The dimensionality reduction has two-fold effects: the noises introduced by the samples or the clustering can be eliminated; on the other hand, the information will lose when the dimensionality is reduced.

4.3 Audio feature

In this system, we use three audio attributes: Rhythm strength, average onset frequency and tempo [15], which are extracted at the shot level. For every one of them, we extract their mean and variance on all the shots to achieve the audio feature in the whole video level. The dimensionality of the resultant feature is 9. It is supposed that the mean reflects the average rhythm for the whole video and the variance represents whether the rhythms often change temporally.

Rhythm information is characteristic for various video categories. Such as, Game videos are often with strong rhythm and quick tempo; while the rhythms of animal videos are often slow and weak. Besides, for the Auto & Vehicle category, the rhythm varies in the time line: slow and weak when the auto stops, but strong and rapid when the auto is running.

Algorithm 2 Manifold *tf-idf* propagation

1: For every word term k_i set $W(i,i) = 0$ and $W(i,j) = 0$ if k_j is not the N nearest neighbors of k_i. Construct the matrix $S = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$, where D is a diagonal matrix with $D_{ii} = \sum_j W_{ij}$.
2: For the video, compute the vector model V_i with *tf-idf* as the weight. $V_i = [t_1, t_2, \ldots, t_k]'$. Set $Y = [V_1, V_2, \ldots, V_n]$.
3: Propagate the *tf-idf* for every word terms. repeat $F_{t+1} = \alpha S F_t + (1-\alpha)Y$ until $F_{t+1} = F_t$. where α is one factor in $(0,1)$.
4: For the i^{th} video, the propagated vector model for the surrounding text is the i^{th} column of the convergent F_t.

4.4 Surrounding text features

One difference between web videos and traditional videos is that web videos are enriched with social information via user interaction on the web. This can be utilized to investigate the relations between videos. In this paper, only associated surrounding text information including title, description and tags are used for categorization.

Information retrieval field has developed many models to model the textual information, including vector model, probability model, and etc. [2]. Given a word dictionary containing n words $(k_1, k_2, \ldots, k_n)$, the vector model representation of a text document $\mathbf{d}$ is: $(w_1, w_2, \ldots, w_n)$ with

$$w_i = freq(k_i, \mathbf{d}) \cdot \log \frac{N}{n_i} \quad (1)$$

where w_i is the so-called *tf-idf* measure.

From our web video corpus collected from the web, consisting of 11333 videos, it is observed that average number of tags is 4.35 and the maximum is 24 for one video. This will result in the sparseness of the vector model. Given the word dictionary containing 1000 (the actual number will be much greater than this) items, the vector model would have at most 24/1000 non-zero entries!

In the other hand, the relationship between the word terms should be considered as well. For example, if a video is tagged with "vehicle" and another video is tagged with "auto," it's naturally supposed that these two videos are similar in the sense of categorization. In order to utilize the word relationship as well as to reduce the sparseness of the vector model, one approach named propagated vector model will be introduced in the following.

4.4.1 Propagated vector model

In order to discover the relationship between word terms, the word similarity computed using the JCN word similarity measure on WordNet[1] is employed. Given a word dictionary containing n words $(k_1, k_2, \ldots, k_n)$, a word similarity matrix W is constructed with the entry $W(i,j)$ being the word similarity between terms k_i and k_j computed based on WordNet.

Given the word similarity matrix W, a graph with the word terms as nodes and the word similarities as the weights of the edges is constructed. Then, manifold ranking [29], a graph based method which is originally used for machine learning, is employed to propagate the *tf-idf* (term frequency inverse document frequency) in the vector model of video texts to similar words. We call it manifold *tf-idf* propagation. The detailed algorithm is depicted in Algorithm 2.

If the *tf-idf* measure is regarded as the similarity between the video and the word, manifold *tf-idf* propagation is to propagate the similarity between the video and some term to other related word terms. Such propagation will give the larger similarity between the videos which have different but related words, such as "music" and "song." Through other contextual word similarity computation, even the video which has tag like "Mariah Carey" and another video which has tag like "music" can be propagated to be more similar.

[1]available at http://search.cpan.org/~tpederse/WordNet-Similarity-0.12/

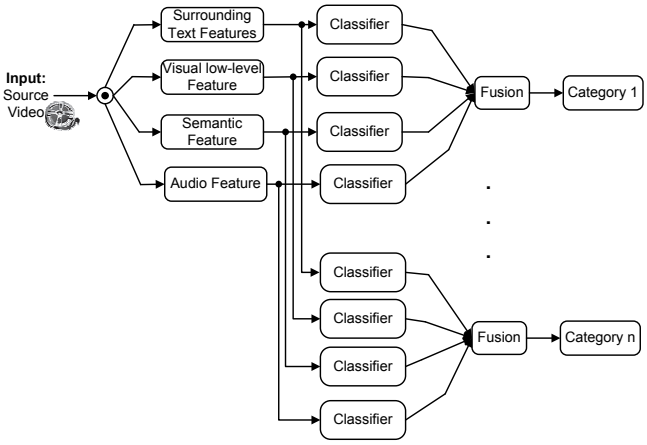

Figure 3: System architecture of Web video categorization. For the input web videos, three steps are processed: (1) Feature extraction; (2) Classification; (3) Fusion.

5. SYSTEM

For the input web videos, three steps are processed. (1) Feature extraction. The features for the four modalities are extracted for each input video. (2) Classification. Each combination of the category and the feature needs a specified classifier. For example, given five features and 11 categories, there will be 55 classifiers in total. Each classifier should output a confidence of whether the input video belongs to the given category based on the given feature. (3) Fusion. For each category, the outputs of the classifiers for different features are fused to achieve a final confidence of whether the video should be annotated this category. For all the input videos, ranking lists are formed based on the confidence for every individual category. Figure 3 is the system framework for web video categorization.

In the classification part, we have experimented 3 kinds of classifiers, including SVM which is a supervised discriminative classifier, GMM which is a supervised generative method and Manifold Ranking which utilizes both labeled and unlabeled data and called semi-supervised learning algorithm. They are briefly described below.

5.1 Support Vector Machine

SVM is a well known machine learning method so that it is commonly employed as a baseline in TRECVID [24]. In video categorization, each category has employed a SVM classifier to output a confidence of whether the input video belong to it based on a given modality. The output of SVM

Algorithm 3 Manifold Ranking

1: Construct the adjacency matrix W for every labeled and unlabeled sample. $W_{ij} = exp(-\|x_i - x_j\|^2/2\delta^2)$ if $i \neq j$; $W_{ij} = 0$ if $i = j$.
2: Construct the matrix $S = D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$, where D is a diagonal matrix with $D_{ii} = \sum_j W_{ij}$.
3: Iterate $F_{t+1} = \alpha S F_t + (1-\alpha)Y$ until $F_{t+1} = F_t$, where Y is a vector with Y_i is the corresponding label if the sample x_i is labeled, or 0 if x_i is unlabeled. α is one factor in $(0, 1)$.
4: The output of all unlabeled samples in F is the resulted rank list.

(here we used LibSVM [4]) is ranked for all the input testing videos. RBF kernel is used. The parameter C and γ are selected based on cross validation. Another issue for adopting SVM in web video categorization is the sample imbalance in positive and negative samples. Here we solve this problem by using different penalty factor C for positive and negative samples. They are chosen by the following criteria:

$$\begin{aligned} W^+ N^+ &= W^- N^- \\ W^+ N^+ + W^- N^- &= N \\ C^+ &= C W^+ \\ C^- &= C W^- \end{aligned} \tag{2}$$

Where C^+ and C^- denote different penalty factor for positive class and negative class, are weighted C with the weights are W^+ and W^-; N^+ and N^- denote the number of positive and negative samples; N is the total number of samples; and $N = N^+ + N^-$. The above equation can be solved:

$$C^+ = C\frac{N}{2N^+} \qquad C^- = C\frac{N}{2N^-} \tag{3}$$

5.2 Gaussian Mixture Model

Gaussian mixture model is a generative classification method, which models the distribution of every class using a mixture of multiple Gaussian models. EM algorithm is normally employed for training. In the experiment, we adopt a diagonal GMM implementation in Torch Library [6]. The optimal number of Gaussian models for each category is selected by cross validation.

5.3 Manifold Ranking

Different from the above two methods, manifold ranking [29] is a semi-supervised method, which learns from the unlabeled data as well as labeled data. The algorithm is described in Algorithm 3. The key idea of this method is to iteratively propagate the label of one sample to its similar samples until a global stable state is reached.

6. EXPERIMENTAL RESULTS

In our experiments, we first compared different feature representations of the four modalities including visual low-level, semantic, audio and text modalities to investigate the best feature representation for each one. We further compared three classification paradigms on the four modalities, including a discriminative classifier (SVM), a generative classifier (GMM) and a semi-supervised method (MR). Finally the fusion experiments of the classifiers based on different features are conducted.

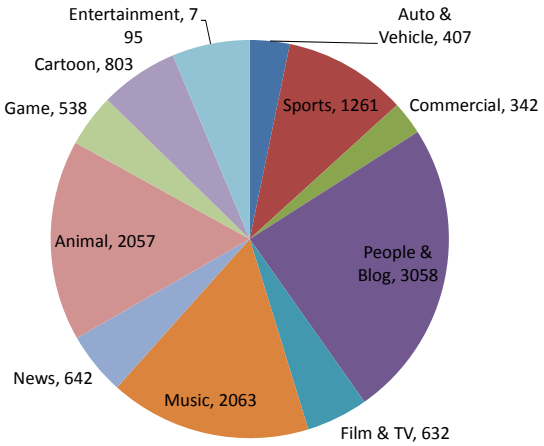

Figure 4: The number of videos for each category

6.1 Dataset

We collected more than $10k$ micro videos from the web, with the surrounding text extracted via parsing the associated web pages. We invited 10 individuals to annotate these video as 11 categories. The annotators were asked to annotate each video with zero, one or more categories at once. The videos which are not annotated any category are eliminated from the dataset. As a result, the dataset is composed of 11333 videos with 300,975 shots and 448 hours in total. The number of videos in each category is shown in Figure 4. In the experiment, the dataset is divided into four parts: 5/10 for training, 3/10 for testing, 1/10 for classifier validation, and the remaining 1/10 for fusion validation.

6.2 Evaluation Measure

In the experiments, we use AP (average precision) as the evaluation measure, which is widely used as a measure of retrieval effectiveness [13] [23]. AP is defined as an average of precisions at various levels of recall, given as

$$AP = \frac{1}{n^+} \sum_{i=1}^{n^+} Precision_i \tag{4}$$

where n is the total number of the samples and n^+ is the number of the positive samples. $Precision_i$ is the precision when i positive samples are returned.

$$Precision_i = \frac{i}{R_i} \tag{5}$$

where R_i is the number of returned samples when i positive are returned.

The AP results among all categories are averaged as mean average precision (MAP).

6.3 Comparisons of different representations for each modality

Figure 5 gives a comparison of the performances by visual low-level features, using SVM as the classifier. It can be seen that the proposed new 17-dimensional visual low-level features outperforms the old 9-dimensional. We can conclude that the new introduced eight dimensions improve the discriminative ability in SVM.

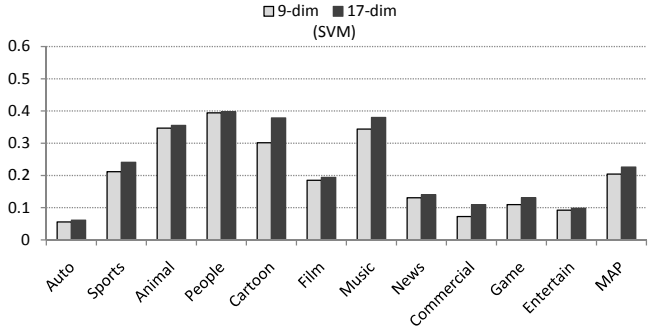

Figure 5: Comparison between the old 9-D visual low-level features and proposed 17-D features using SVM.

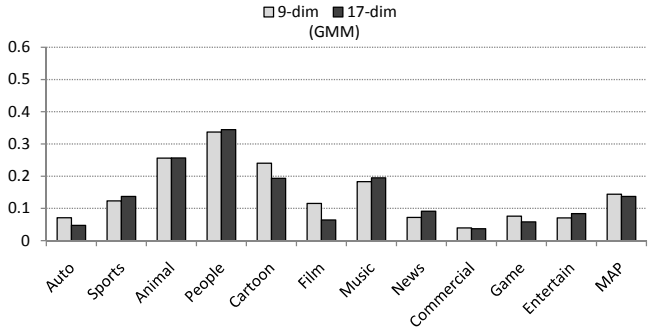

Figure 6: Comparison between the old 9-D visual low-level features and proposed 17-D features using GMM. It shows that the new introduced features decrease the modeling ability.

However, these eight dimensions also introduce some noises in the samples. As shown in Figure 6, these features lead GMM to performance degradation. We can conclude that the new features decrease the modeling ability in GMM.

For semantic modalities, we compared the representation by using visual word vector model, visual word LSA and concept histogram. Experimental results in Figure 7 show that visual word LSA outperforms visual word vector model and concept histogram. Hence, visual word LSA (the number of clusters is 400) is selected as the feature representation for semantic modality.

Figure 8 shows the comparison between visual low-level feature and semantic feature. We can see that the MAP of semantic feature is higher than that of low-level feature. This is because that web video categories are more about content. It is also observed that for genre-related categories such as sports, news and commercial, visual low-level feature outperforms semantic feature. This proves that visual low-level feature is more suitable for genre-related categories than semantic feature, vise visa for content-related categories such as people, auto, and so on. Exceptions exist for cartoon, film, and music. As explained in Section 3.2, music category in this context is not the same as the definition of genre "music video." For Film (referred to as the category "Film & TV" in this paper), as described in section 3.2, since its content usually has large variation due to compression, editing and direct capturing of TV screen, the visual low-level feature may not work well. By observing cartoon videos it is reasonable that semantic feature via

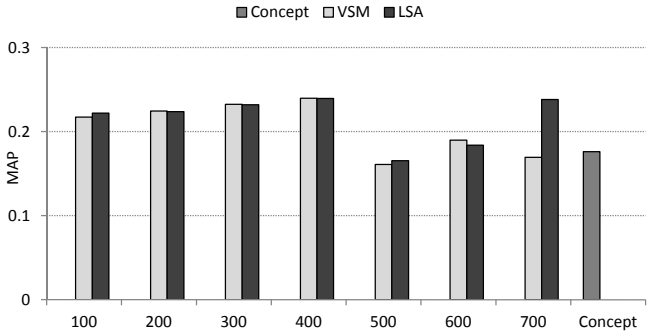

Figure 7: The comparison among three semantic features: concept histogram (Concept), visual word vector model (VSM) and visual word LSA (LSA) using SVM as the classifier. The horizontal axis is the number of clusters of the visual words.

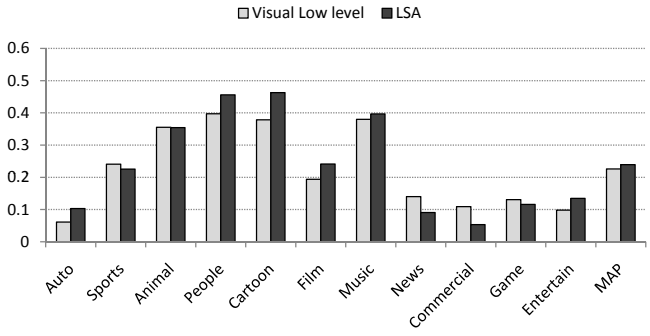

Figure 8: Comparison between visual low-level feature and semantic feature. The visual low-level is the new proposed 17-D feature and the semantic feature is the visual word LSA with 400 clusters. The classifier is SVM.

automatic visual word construction outperforms visual low-level feature.

For surrounding text, experiments are conducted for title (referred to as title and description here) and tags, respectively. As some videos collected from the web have no textual information, the experimental test set is only a subset, i.e., 2120 samples, compared with the total test dataset containing 3394 videos. We call it *reduced test set*. Figure 9 gives the experimental results of classic vector model and propagated vector model for the surrounding text. The MAP of propagated vector model for title is a little better than the classic. The APs of some categories are better, and others are worse. This is because that in the surrounding text there are many noises and the noises will be propagated as well when manifold *tf-idf* propagation is applied. In the future work web text de-noising should be considered to achieve a better propagation result. For tags, propagation will decrease the performance. From the observation on the tags, we can conclude that it is because that the tags are nosier than title and description. For example, the video in "People" category may have the tags like "sing" and "soccer." However, the video does not belong to music or sports category. The propagated vector model should perform better if only the tags are pre-processed and de-noised. Therefore, for the text modality, we propose to use two features,

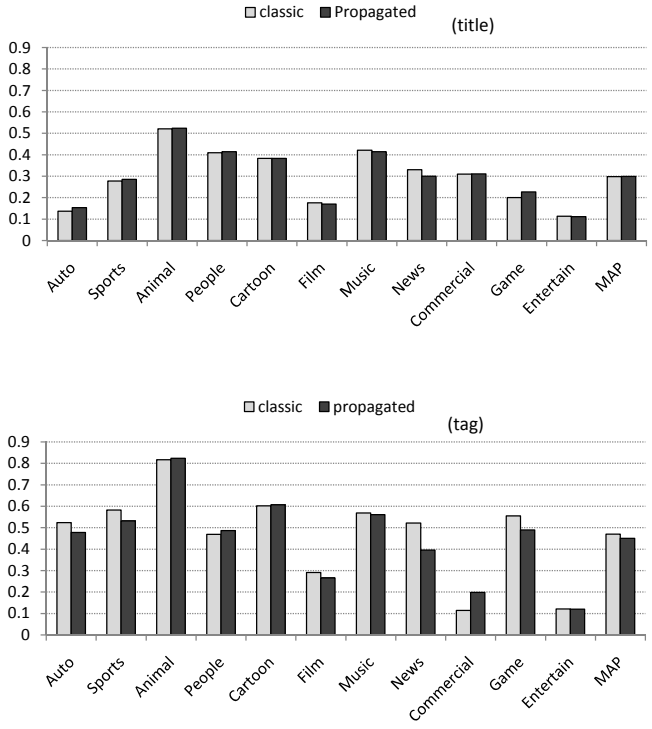

Figure 9: Comparison between classic vector model and propagated vector model for the surrounding text using SVM as the classifier. The above is for title and description, and below is for tags.

i.e., title propagated vector model and tag vector model, as the representation.

6.4 Comparisons of classifiers

We compared the performances of the three classifiers (i.e., SVM, GMM, and MR) based on the selected features in Section 6.3. (i.e., visual low-level, visual word LSA, audio, title propagated vector model and tag vector model). From figure 10, we can observe that SVM significantly outperforms the other two classifiers. We argue that the ineffectiveness of GMM is probably because the features are extremely varying for each category and cannot be generated by some parametric distribution such as GMM. In addition, because of the large intra-class variance, the prior consistency assumption [29] of MR cannot be satisfied.

6.5 Fusion

First we conducted the fusion experiment on the *reduced test set* consisting of 2120 test videos. We used the *max* fusion, *average* fusion, and *linear weighted* fusion [18] (the weight is selected according to the AP of each single modality) on the SVM outputs. As shown in figure 11 the *linear weighted* fusion outperforms the other two fusion methods. In addition, *average* fusion was also experimented on all the outputs of the three classifiers. The results show that ineffective classifiers decrease the performance.

Second we conducted fusion experiment on the whole test dataset. If one sample has some missing features such as textual information, this feature was not used. The results are shown in Figure 12. It can be observed that the *linear weighted* fusion (weighted as the above) outperforms the other two. In order to show the performance of our system on the web videos without textual information, one experiment which only fused the visual low-level, audio and semantic features was conducted. Figure 12 shows the results. It can be observed that the fusion outperforms all of the single modality.

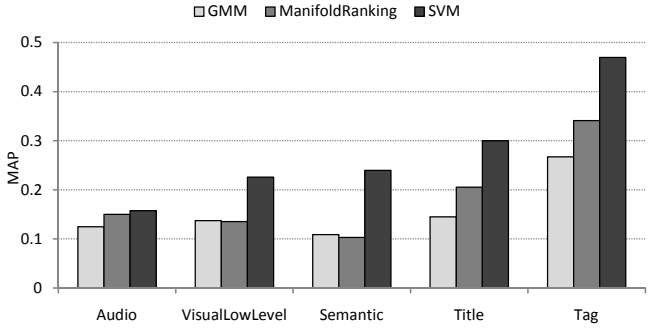

Figure 10: Comparisons of three classifiers: SVM, GMM and Manifold Ranking.

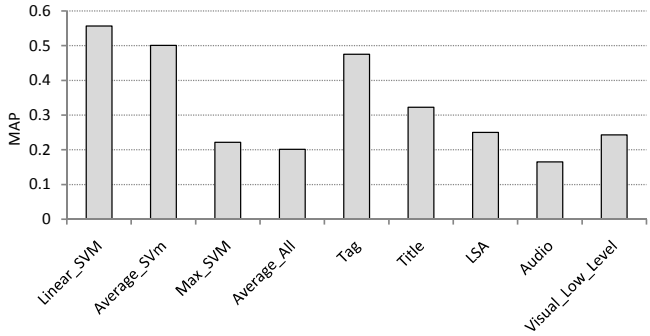

Figure 11: The fusion results on the *reduced test set*. Linear_SVM indicates linear combination of SVM outputs. Average_SVM and Max_SVM are average and max of all outputs of SVM. Average_All combines all the outputs of the three classifiers with various modalities. The columns of "Tag," "Title," "LSA," "Audio," and "Visual_Low_Level" in the figure indicate the performances of every single modality on the *reduced test set*.

7. CONCLUSIONS AND FUTURE WORK

This paper conducted a comprehensive study on web video categorization. From the observation that the quality, subject, style, and genres of web videos are extremely diverse, novel modalities including semantic and surrounding text are proposed to complement the low-level features which are commonly used in existing video genre classification. Furthermore, we designed a multi-modality web video categorization system with selected feature representations for different modalities. The experiments over a relatively large dataset demonstrated that MAP of 0.56 was achieved by *linear weighted* fusion. We believe that the comprehensive analysis and tests of multi-modality features presented here are relevant and useful to researchers and practitioners for developing related technologies on web videos.

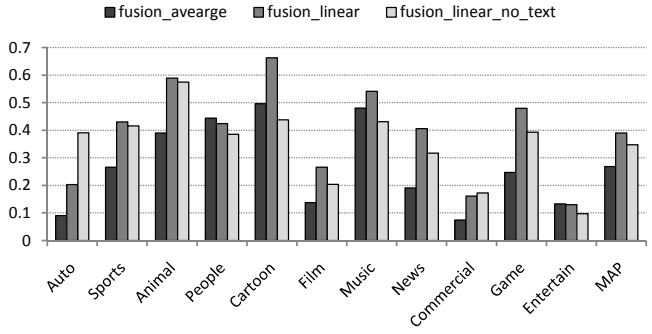

Figure 12: Fusion on the whole test dataset. Average fusion and linear fusion are experimented. Considering some videos without texts, linear weighted fusion of the modalities except text is also experimented.

Based on the study of web video categorization, it can be concluded that: (1) The newly-proposed two modalities are effective on web video categorization. Semantic modality performs better than visual low-level features for content related categories. Surrounding text outperforms all the other modalities. (2) For web video categorization, supervised discriminative classifier (SVM) is much better than generative (GMM) and semi-supervised (MR). (3) The multi-modality representation is superior to each individual.

In the future work, we will work on the following problems:

- Multi-label learning algorithms should be taken into account for web video categorization, since each web video may have multiple categories and some of the categories exhibit correlative relationship.
- Surrounding text de-noising. As described in the experiment, the noises in the texts decrease the performance of propagated vector model. If the noises are removed, such propagation may achieve a better performance.
- The ontology for web videos should be defined not only from the perspective of video categorization, but also semantic video annotation. We believe this will help improve video categorization.

8. REFERENCES

[1] 6rooms. http://www.6rooms.com/.

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[3] M. Campbell, S. Ebadollahi, D. Joshi, M. Naphade, A. Natsev, J. Seidl, J. R. Smith, K. Scheinberg, J. Tesic, L. Xie, and A. Haubold. IBM Research TRECVID-2006 Video Retrieval System. In *TREC Video Retrieval Evaluation Online Proceedings*, 2006.

[4] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines.

[5] S.-F. Chang, Winston, W. Jiang, L. Kennedy, D. Xu, A. Yanagawa, and E. Zavesky. Columbia University TRECVID-2006 Video Search and High-Level Feature Extraction. In *TREC Video Retrieval Evaluation Online Proceedings*, 2006.

[6] R. Collobert, S. Bengio, and J. Marithoz. Torch Lib for GMM and EM.

[7] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[8] S. Fischer, R. Lienhart, and W. Effelsberg. Automatic recogition of film genres. In *Proceedings of ACM Multimedia*, pages 295–304, San Francisco, CA, Nov. 1995.

[9] Google Web Directory. http://directory.google.com/.

[10] A. G. Hauptmann, M.-Y. Chen, M. Christel, W.-H. Lin, R. Yan, and J. Yang. Multi-Lingual Broadcast News Retrieval. In *TREC Video Retrieval Evaluation Online Proceedings*, 2006.

[11] X.-S. Hua, T. Mei, W. Lai, and *et al.* Microsoft Research Asia TRECVID 2006 high-level feature extraction and rushes exploitation. In *TREC Video Retrieval Evaluation Online Proceedings*, 2006.

[12] Jumpcut. http://www.jumpcut.com/.

[13] K. Kishida. Property of average precision and its generalization: An examination of evaluation indicator for information retrieval experiments. *Nii technical report (nii-2005-014e)*, 2005.

[14] R. Lienhart. Abstracting home video automatically. In *Proceedings of ACM International Conference on Multimedia*, pages 37–40, Orlando, FL, USA, Oct 1999.

[15] L. Lu, D. Liu, and H.-J. Zhang. Automatic mood detection and tracking of music audio signals. *Special Issue on Statistical and Perceptual Audio Processing, IEEE Trans. on Audio, Speech and Language Processing*, 14(1):5–18, 2006.

[16] MSN Soapbox. http://soapbox.msn.com/.

[17] Online Publishers. http://www.online-publishers.org/.

[18] R. Ranawana and V. Palade. Multi-classifier systems: Review and a roadmap for developers. *International Journal of Hybrid Intelligent Systems*, 2006.

[19] Z. Rasheed, Y. Sheikh, and M. Shah. On the use of computable features for film classification. *IEEE Trans. on Circuit and System for Video Technology*, 15(1):52–64, Jan 2005.

[20] C. G. M. Snoek, M. Worring, J. C. van Gemert, J. M. Geusebroek, and A. W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the ACM International Conference on Multimedia*, Santa Barbara, USA, 2006.

[21] F. Souvannavong, B. Merialdo, and B. Huet. Latent semantic analysis for an effective region-based video shot retrieval system. In *Proceedings of ACM SIGMM International Workshop on Multimedia Information Retrieval*, October 2004.

[22] TechWeb. http://www.techweb.com/.

[23] TRECMeasure. http://trec.nist.gov/pubs/trec10/appendices/measures.pdf.

[24] TRECVID. http://www-nlpir.nist.gov/projects/trecvid/.

[25] B.-T. Truong, S. Venkatesh, and C. Dorai. Automatic genre identification for content-based video categorization. In *Proceedings of ICPR*, pages 1–10, 2000.

[26] Yahoo! Video. http://video.yahoo.com/.

[27] YouTube. http://www.youtube.com/.

[28] X. Yuan, W. Lai, T. Mei, X.-S. Hua, and X.-Q. Wu. Automatic video genre categorization using hierarchical svm. In *Proceedings of IEEE International Conference on Image Processing*, Atlanta, USA, Oct. 2006.

[29] D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Scholkopf. Learning with local and global consistency. In *Proceedings of Advances in Neural Information Processing System*, 2004.