Features with Feelings—Incorporating User Preferences in Video Categorization

Ramya Srinivasan and Amit K Roy-Chowdhury

Electrical Engineering Department, University of California, Riverside, 92521

Abstract. Rapid growth of video content over internet has necessitated an immediate need to organize these large databases into meaningful categories. In this paper, we explore the benefits of leveraging social attitudes (beliefs, opinions, interests and evaluations of people) with machine learning concepts (audio/video features) in the challenging and pressing task of organization of online video databases. Through the analysis of view counts, we model social participation (people's choices) towards a video's contents. Observations reveal that viewership patterns are correlated with video genres. We propose logistic growth models to characterize videos based on usage and obtain a probability of video category. We then combine these subjectively assessed priors with likelihood of video class (as estimated from objective audio/video features) to establish the final category in a Bayesian framework. We provide a comparitive analysis of classification accuracies when a) categories are known a priori b) when they are not known a priori. Experimentally, we establish improvement in classification accuracy upon incorporating social attitudes with state-of-the-art audio/video features. [1]

1 Introduction

Due to the ease with which they can be created, video content has been growing at a rapid rate over the internet. In order to provide easy and quick access to useful information and thereby facilitate better user experience, there arises a need to organize these large video databases into meaningful categories. However, this is a challenging problem. The sheer volume of such data imposes several constraints on their organization. Further, due to unbounded diversity of videos both in content and quality, analysis of these videos is more challenging.

Consider for instance the use of tags for categorization. Tags are often associated with personalized opinions biased by culture and region of the user. These could be in different languages and prone to mis-spellings. Moreover such metadata can include homonyms and synonyms which can make their use inefficient. For example "watch" could refer to the the verb meaning "look" or to the noun meaning "clock". Moreover, portals such as Youtube, are subjected to spamming due to which there could be lot of unrelated tags. The flexibility of tagging allows users to classify items in the way they find useful, thus presenting challenges for categorization.

[1] This work was partially supported under NSF grant IIS-0712253

Although more robust than methods based on textual and audio features, methods based on visual features are computationally intense and challenging. Classification results employing visual features largely depend on the effectiveness of features to represent underlying video patterns. More number of features can provide more discriminating power and hence enable better classification. However, time requirements for classifier training grow at a very high rate with feature dimensions. These limitations motivate exploration of new dimensions to tackle the problem.

Since classification is fundamentally an act of defining bounds based on the *perceived* similarities that is common to all members belonging to a group, the role of cognitive elements (beliefs, opinions and evaluations) in predicting relations of synonymy or difference is vital. These cognitive elements fall under the broad concept of social attitudes [5].

Social attitudes can be defined as the tendency of a person or a group of people to behave in a certain way. These include social emotions, reasoning and intuitions, apart from opinions, beliefs and evaluations. All these elements determine preferences and intentions. We believe that people's preferences can provide valuable input in understanding a video's contents and that this additional knowledge can improve classification accuracy. In doing so, we do not want to involve people directly since it is only going to make the process very involved. We would like to benefit from those sources of information which people unknowingly provide and which are readily available.

1.1 User preferences for various video genres

One of the ways of measuring users' preferences is by the number of views a video accumulates over time. Observations reveal that different videos genres generate different viewership patterns. Music videos are the most popular among users. They generate huge interest over a long period of time. Entertainment videos attract user attention for about a week before being overtaken by newer episodes that are uploaded. Although sports videos generate considerable interest, people probably prefer to watch them through more exciting channels such as a live telecast or in the stadium and thus unlike entertainment videos, these videos do not hold people's interest for a long time in the online medium.

Except for the ones concerning major events like natural disasters, most other news videos trigger limited interest amongst a small audience. Certain other videos, mostly personal videos such as cooking, home parties etc fail to capture the attention of considerable viewers even for about a week. Based on these observations, we hypothesize that usage statistics can help in distinguishing categories.

Indeed, we noticed characteristic curves for cumulative viewcounts. Fig. 1 shows the smoothed cumulative viewcount for randomly chosen videos belonging to different categories. As we can notice from the figure, different categories exhibit different growth rates. These curves are somewhat similar to logistic functions. In fact, logistic functions have been used to model biological population growth in species, to characterize spread of information within societies

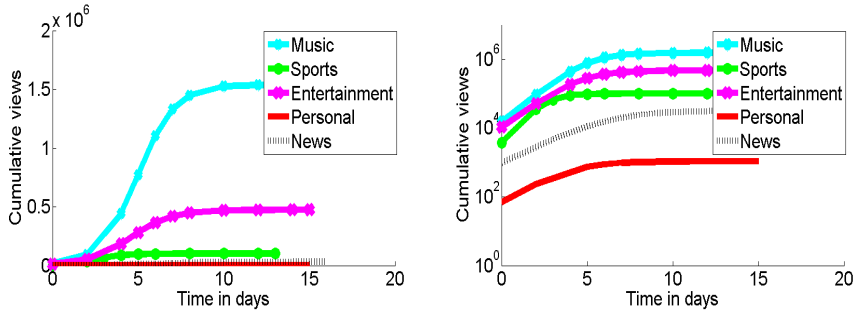

Fig. 1. Figure on left shows cumulative viewcounts for randomly chosen individual videos belonging to different categories. X axis denotes time in days and y axis denotes cumulative views. For clarity, figure on right shows same information on y axis in log scale.

and in marketing to model sales of new products among others [9]. *We model viewcounts in video databases using logistic functions.*

It should be noted that our goal is to explore the role of usage statistics towards improved video classification. Several complex factors characterize growth of views, video genre being one of them. While it is true that a video could be mis-classified based on usage alone (just as the case could be using audio-video features alone), it is to be noted that we are using usage statistics as just a prior.

We are thus not claiming that usage statistics can reveal the true category by itself, but would like to show that when integrated with state-of-the-art audio/video features, it can provide considerable improvement in classification accuracies. It is often the case that complementary sources of information increase classifcation accuracy and usage statistics is complementary to audio-video features.

1.2 Contributions of the work

We propose a novel method for robust web video classification by fusing usage statistics with low level audio-video features. The proposed approach is independent of cultural and regional variances commonly associated with tags and is computationally efficient in handling large databases where using just visual features can be involving.

We consider the scenarios when categories are known apriori and not known apriori and show that in both scenarios, combining usage information with audio-video features yields considerable improvement in classification accuracy. Towards this, we make the following contributions.

1. Propose a logistic growth model to characterize videos based on their usage;
2. Estimate probability of video category based on the above mentioned characteristics ; and
3. Integrate audio/visual features with usage statistics through a Bayesian framework to establish final video categories.

1.3 Related work

There are a number of papers which address the problem of automatic video classification. A good survey of these papers is provided in [1].

More recent works in this area have focused on improving classification accuracy using text metadata, related videos, user comments and other social information. A semantic feature space with high distinguishing ability to represent web videos was proposed in [7]. In [4], it was shown that two complementary views on the data from text and video perspectives refine predictions of video category. A statistical approach built upon local and global features for classification of consumer videos was considered in [2].

In [6], the authors proposed a classification system that exploits co-browsed, co-uploaded, co-commented and co-queried videos to extract correlated information. In [8], a web text document trained classifier was applied to the video domain so as to utilize the large set of labeled text documents. In the recent past, contextual models have been explored to classify videos. Authors in [10] used semantic meaning of text, related videos and videos uploaded by the same user to determine video category.

However, most of the previous works relied on textual features which are sparse, prone to mis-spellings or which may be in different languages. Our method can consider inputs from different societies without needing language recognition for each. Further, we propose to use *passive metadata* (which does not involve active participation of people in the form of comments/ratings) like usage statistics as a context for classifying videos.

One another method that has been widely explored for image/video retrieval and classification is using relevance feedback from users. A good survey of the applications of user feedback for information retrieval is provided in [11]. Since this aspect has been studied widely, it is not a focus of our work.

Authors in [3] make use of usage statistics to argue that videos appearing on front page have different history than other videos and draw analogy from book sales to categorize. Although we explore usage statistics for classification, unlike [3], we analyze social dynamics of videos by understanding the growth of their views through a logistic model to classify into a diverse set of classes.

1.4 Overview of the approach

The overall scheme of the classification process is depicted in Fig. 2. Unorganized (test) videos are the inputs to the system. Analysis of view trends of these videos helps in obtaining probability of video belonging to a certain class. Towards this, we employ logistic function to model growth of view counts. We use these logistic parameters as feature vectors to discover category clusters (Kmeans algorithm) when the classes are unknown a priori to obtain probability of a video's class. When classes are known a priori, we use the average values of these logistic parameters of each class to predict the probabilty of a video's class in a mixture model framework. Thereafter, we integrate likelihoods of audio/video features

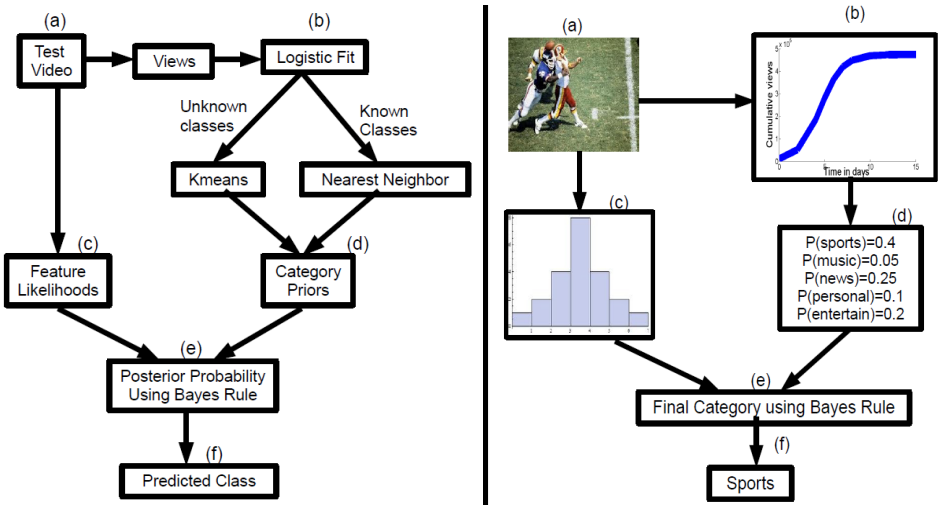

Fig. 2. Figure on left provides the general methodology of the approach. A particular illustration of the approach on a random test video (sports in this case) is provided on the right. Viewcounts of test video are tabulated at a regular basis (a) and a logistic function is fitted to it in the least squared error sense (b).Probability of video category is estimated based on views (d). This is integrated with audio video feature likelihood (c) in a Bayesian framework (e) to predict the aposteriori class (f).

with usage priors in a Bayesian framework to obtain the maximum a poteriori class of test videos.

The rest of the paper is organized as follows. Section 2 discusses the model of usage statistics. Section 3 describes the objective features we have used and presents the integration framework through which we incorporate usage statistics with audio/visual features. Experimental results are discussed in Section 4. Section 5 presents conclusions.

2 Modeling usage statistics

Phenomena such as beliefs, opinions, etc., which constitute social attitudes can be primararily thought of as complex, well co-ordinated sources of information which are characteristic of the interplay between individuals and groups. Individual and social systems co-evolve such that individual systems reflect the chracteristics of social systems—their beliefs and opinions; while social systems reflect characteristics of individual systems—their temporal variations etc.

The co-evolutionary description of development between social and individual systems explains why, and to some extent how, collective opinions of users formed by the aggregation of individual opinions can benefit video classification. Thus we analyze the patterns of viewcount growth for different video genres to understand the underlying preferences of users. Towards this, we propose a logistic model which we discuss next.

2.1 Logistic model fit

The general formula for the logistic function can be written as

$$y = \frac{C}{1 + Ae^{-Bt}}. \tag{1}$$

The parameter C represents the limiting value of the population (often known as carrying capacity). A can be interpreted as the number of times the initial population must grow to reach C. Parameter B decides the way a logistic function behaves—positive B indicates increasing logistic function and negative B indicates decreasing logistic function.

Viewcounts are tabulated on a regular basis for all videos. We then fit a logistic function in the least squared error sense to these viewcounts for every video. The individual $[A, B, C]$ parameters characterize each video.

2.2 Prior estimate of video category when classes are unknown

First we consider the case when the categories to which videos belong are unknown. Let n be the total number of videos belonging to all categories in the database. $\boldsymbol{x} = [A, B, C]$, the logistic parameters for individual videos form the observation space. Kmeans algorithm is used to partition the n dimensional observation space into k subsets ($k \leq n$) chosen randomly.

Let c_i^t denote the i^{th} cluster at time step t, $\boldsymbol{m}_i^t$ denote the centroid of the i^{th} cluster at time t and x_p denote the observation vector of logistic parameters. The algorithm then proceeds by alternating between the following steps until there is no significant change in the clusters to which individual videos belong.

1. $c_i^t = \boldsymbol{x}_p : ||\boldsymbol{x}_p - \boldsymbol{m}_i^t|| \leq \left|\left|\boldsymbol{x}_p - \boldsymbol{m}_j^t\right|\right| \forall 1 \leq j \leq k$. (i.e., assign each observation to its closest cluster centroid)

2. $\boldsymbol{m}_i^{t+1} = \frac{1}{|c_i^t|} \sum_{\boldsymbol{x}_j \in c_i^t} \boldsymbol{x}_j$ (i.e., replace the cluster centroids)

We then obtain P_{c_i}, the probability with which a video can belong to cluster c_i as $P_{c_i} \propto \frac{1}{d_{vi}}$ where d_{vi} is the normalized Euclidean distance between the video and the corresponding cluster centroid. We integrate this with audio-video features as discussed in Sec 3.

2.3 Prior estimate of video category when classes are known

We now consider the case when categories are known a priori. We use average values of A,B,C parameters obtained for different known categories in a mixture model framework to estimate probabilities. We do this by computing the average values of logistic parameters across all training videos belonging to a category. Let these parameters be denoted by $A_{c_i}, B_{c_i}, C_{c_i}$ for category c_i. Let

$$\boldsymbol{b} = \begin{pmatrix} A_{c_1} & B_{c_1} & C_{c_1} \\ A_{c_2} & B_{c_2} & C_{c_2} \\ \cdots & & \\ A_{c_k} & B_{c_k} & C_{c_k} \end{pmatrix}$$

where there are k categories, $\boldsymbol{a} = [A_t, B_t, C_t]$ be the test video logistic parameters and $\boldsymbol{P} = [p_1, p_2, \cdots, p_k]$ where p_i is the probability with which a test video can belong to class c_i. We then seek to express $\boldsymbol{a}$ as

$$\boldsymbol{a} = p_1[A_{c_1}B_{c_1}C_{c_1}] + p_2[A_{c_2}B_{c_2}C_{c_2}] + \cdots p_k[A_{c_k}B_{c_k}C_{c_k}] \tag{2}$$

The solutions of the above equation is equivalent to computing the following.

$$\min_{\boldsymbol{P}} ||\boldsymbol{a} - \boldsymbol{P}\boldsymbol{b}||_2^2 \quad s.t \quad ||\boldsymbol{P}||_1 = 1, p_i \geq 0 \tag{3}$$

The above is a constrained optimization problem which can be solved by applying Karush-Kuhn-Tucker (KKT) conditions i.e.

$$D\,||\boldsymbol{a} - \boldsymbol{P}\boldsymbol{b}||_2^2 + \lambda D(\boldsymbol{P}.\boldsymbol{1} - 1) - \boldsymbol{\mu}\boldsymbol{P}^T = 0 \tag{4}$$

where D stands for derivative, $\boldsymbol{1}$ is a unit column vector with k rows, λ is the Lagrange multiplier and $\boldsymbol{\mu} = [\mu_1, \mu_2 \cdots \mu_k]$ is the KKT multiplier such that $\mu_i \geq 0$ for all i and $\boldsymbol{\mu}\boldsymbol{P}^T = 0$. Thus we get the probabilities with which a test video can belong to individual categories.

3 Integration of usage statistics with audio-video features

3.1 Audio-video features

We use color histograms and histogram of optical flow for the video feature, and audio energy and frequency bandwidth of the audio signal in our analysis. A number of works employ similar features for classification [1]. Color histograms capture the distribution of colors in an image which vary from genre to genre. Optical flow estimates motion across video frames and audio features estimates volume of sound.

We concatenate these color, motion and audio histograms to obtain a single feature histogram H_t, i.e., a 34 bin histogram with the first 24 bins corresponding to color features, followed by 8 bin optical flow features and last two bins denoting audio features. We also compute average value of these feature histograms of all videos belonging to a category to obtain category average values H_{c_i}. A video is assigned to a category c_i^* as

$$c_i^* = \arg\min d_{ct} \tag{5}$$

where d_{ct} is the Bhattacharya distance between H_t and H_{c_i}

3.2 Integration framework

We employ a Bayesian framework to integrate usage statistics priors with audio/video feature histogram. Let $P_{c_i}(T = c_i)$ denote the (prior) probability with which a video T can belong to a class c_i based on usage statistics as described in Sec 2.2 or Sec 2.3 depending on whether classes are unknown or known a priori.

We compute the likelihood $P_l(D_t|C = c_i)$ based on d_{ct} as defined in (5). We normalize these distances and assign the individual likelihood probabilities on a uniform scale based on the distances between the video and corresponding category.

The (aposteriori) probability of the video to belong to class c_i given its feature histogram D_t is given by,

$$P_t(T = c_i|D_t) = \frac{P_l(D_t|C = c_i)P_{c_i}(T = c_i)}{\sum_{i=1}^{k} P_l(D_t|T = c_i)P_{c_i}(T = c_i)} \tag{6}$$

where there are k classes.

4 Experiments

4.1 Description of dataset

We have evaluated performance of the approach on our dataset that consists of over one thousand videos belonging to diverse categories such as sports, news, entertainment shows, personal videos and music videos. Due to the non-availability of any standard dataset which provides view statistics, we randomly collected videos just after they were uploaded on Youtube to keep track of their view-counts on a regular basis. On the whole, the dataset comprises of 235 music videos, 336 entertainment videos, 238 sports videos, 256 news videos and 108 personal videos spanning over 2000 min. A sample representation of the dataset is shown in Fig 3. The dataset will be available on our website.

Fig. 3. Sample representation of the dataset.

4.2 Predicting probability of video category from usage

In this part of the experiment, our goal was to determine probability of a test video category based on its view trends so as to be able to integrate this information with audio/video features before finally establishing the category.

Logistic parameter fit: We learned the values of the logistic parameters of all videos. The mean and standard deviations of individual parameters for each class when classes are known and unknown a priori is provided in Table 1. Based on these values, we fitted the logistic function for each class.

With unknown classes: We estimated the probability of video category as discussed in Section 2.2. We heuristically varied k, the number of clusters

Category	A		B		C	
	Known	Unknown	Known	Unknown	Known	Unknown
Music	$\mu : 99.2$ $\sigma : 3.2$	$\mu : 97.4$ $\sigma : 4.8$	$\mu : 0.9$ $\sigma : 0.05$	$\mu : 0.9$ $\sigma : 0.19$	$\mu : 3.2E6$ $\sigma : 3891$	$\mu : 2.9E6$ $\sigma : 5100$
Entertainment	$\mu : 43.4$ $\sigma : 0.5$	$\mu : 52.8$ $\sigma : 0.6$	$\mu : 0.8$ $\sigma : 0.08$	$\mu : 0.7$ $\sigma : 0.25$	$\mu : 4.2E5$ $\sigma : 972$	$\mu : 5.1E5$ $\sigma : 1875$
Sports	$\mu : 25.6$ $\sigma : 2.3$	$\mu : 23.5$ $\sigma : 4.2$	$\mu : 1.3$ $\sigma : 0.1$	$\mu : 1.2$ $\sigma : 0.3$	$\mu : 9.8E4$ $\sigma : 410$	$\mu : 7.6E4$ $\sigma : 960$
News	$\mu : 32.2$ $\sigma : 1.2$	$\mu : 29.4$ $\sigma : 3.5$	$\mu : 0.5$ $\sigma : 0.1$	$\mu : 0.4$ $\sigma : 0.3$	$\mu : 3.1E4$ $\sigma : 200$	$\mu : 2.6E4$ $\sigma : 800$
Personal	$\mu : 16.7$ $\sigma : 1.3$	$\mu : 12.6$ $\sigma : 2.2$	$\mu : 0.72$ $\sigma : 0.02$	$\mu : 0.61$ $\sigma : 0.13$	$\mu : 1083$ $\sigma : 45$	$\mu : 786$ $\sigma : 165$

Table 1. Mean and standard deviations of the logistic parameters for various classes under the two scenarios: when categories are known apriori and when categories are unknown a priori.

considered by the Kmeans algorithm. These values were established after manual validation of the categories in each cluster. Best results (with an overall precision of 62.4%) was obtained with k=5, thus predicting 5 major categories.

With known classes: Following the procedure described in Section 2.3, we obtained probability of test videos based on their usage. On an average in about 65% of the cases, usage statistics predicted the true categories with high probability. Particularly for music videos, it was as high as 75 % and for personal videos it was close to 86%.

4.3 Integration with audio-video features

As described in Section 3.2, we integrated usage statistics information with audio-video features to establish categories. Upon integrating usage priors with audio-video features, there was improvement in classification results as against results obtained using just audio-video features (as described in Sec 3.1), this improvement being significant for some categories. Fig 4 provides a comparison of the precisions obtained for various categories under different scenarios namely classification using audio-video features only, classification with both usage statistics plus audio/video features when classes are known apriori and classification with both usage statistics plus audio/video features when classes are unknown apriori.

Discussion of results: As can be seen from Fig 4, there was significant improvement in classification accuracies upon using usage information along with audio/video features for all the categories. For music and personal videos, this was close to 15%. This is because viewtrends of these videos were markedly different from other videos and thus usage could identify a clear distinction from other categories. However, this improvement was not as pronounced in the case of sports, news and entertainment videos. This is because viewtrends of these videos vary widely and thus usage statistics could give erroneous results in some

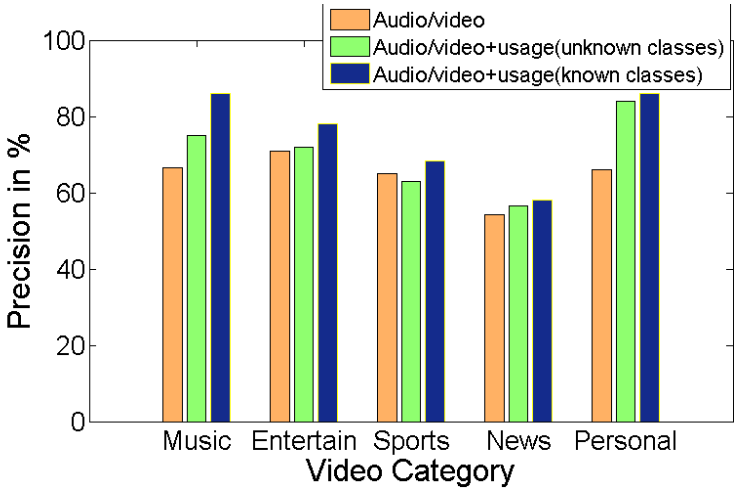

Fig. 4. Classification accuracies for various categories.

cases. When categories are known apriori, better accuracies are obtained.
We also show precision recall curves for various categories for the two scenarios in Fig 5. One can see improvement upon incorporating usage with audio-video features. Overall analysis of these curves showed that initially, both precision and recall increased since the retrieved videos were relevant. Subsequently, as some non-relevant videos were retrieved, recall increased but precision decreased. Fairly high precision values can be obtained even at a higher recall rate when usage statistics is combined with audio-video features; this improvement being significant for all classes considered.

Analysis of the overall approach through case studies: We provide an analysis of the overall system through some case studies. In particular, we address (a) when does usage statistics help/not help? (b)when do audio-video features help/not help? (c)how can fusing these multi-modal features offer benefits?
Case 1: A not so popular music video : This music video gathered only a few hundreds of views during the observation time. As a result, the system predicted it as a personal video with probability 0.7 and estimated its probability of being a music video to be as less as 0.08. Analysis of its audio-video features alone classified it as music video. Combining both features classified it as personal video. Thus, if a video is very rarely seen it is possible that this system can misclassify.
Case 2: A news broadcast: This video featured several events such as a basketball match and war scenes, apart from others. Analysis of audio-video features classified this as a sports video. However, since this video exhibited view trends similar to that of news videos in the database, usage statistics predicted it to be a news video with probability 0.7 while its probability of sports video as 0.15. Combining these features correctly classified this as a news video. Thus in videos with diverse scenes such as that of news, this system can yield better results.

Case 3: Minecraft show: This entertainment video was mis-classified to be a music video when just audio-video features were used. This might have been due to significant audio features in the video. Usage statistics predicted it to be an entertainment video with probability 0.67 and its probability to be a music video to be as less as 0.12. Upon combining both features, it was correctly classiifed.
Case 4: A news video featuring tsunami in Japan: Although this was a news video, it generated lot of interest among the users since it featured a significant event. As a result, usage statistics predicted it to be a music video with probability 0.6. Upon incorporating audio-video features, it was correctly classified as a news video. Fig 6 provides an illustration of the overall performance of the system.

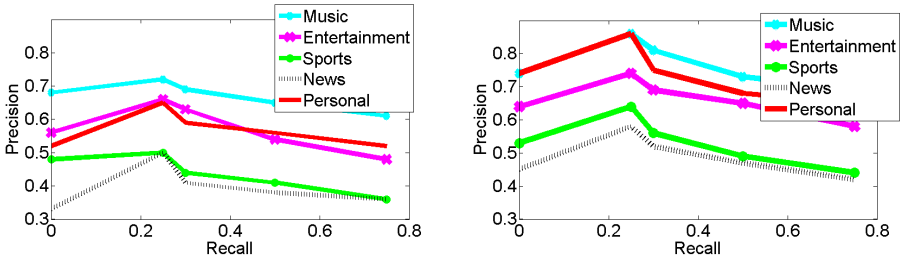

Fig. 5. Precision recall curves for various categories. Figure on left indicates results obtained using only audio-video features while figure on right denotes results obtained upon incorporating usage statistics with audio/video features.

5 Conclusions

We presented an approach that combines subjective perspectives with objective methodologies for improved video classification. Towards this, we proposed a logistic model to characterize videos based on their usage to obtain prior estimate of class. This information was then integrated with audio/video features in a Bayesian framework to establish categories. Results showed improvement in accuracy upon incorporating usage statistics both when classes were known apriori and when they were unknown apriori. The proposed approach makes use of readily available usage statistics information with state-of-the art audio/video features, is invariant to regional and cultural variances prevelant in textual features,is computationally efficient and easily scalable to handle large datasets. As future work, we would like to explore the scalability of the method to larger databases and its applications to web traffic monitoring.

References

1. Brezeale, D., Cook, D.: Automatic Video Classification: A Survey of Literature. IEEE Trans. on Sys. Man and Cybernetics **38(3)** (2008)
2. Chang, S., Ellis, D., Jiang, W., Lee, K., Yanagawa, A., Loui,A., Luo, J.: Large Scale Multimodal Semantic Concept Detection for Consumer Video. Intl. Workshop on Multimedia Information Retrieval (2007)

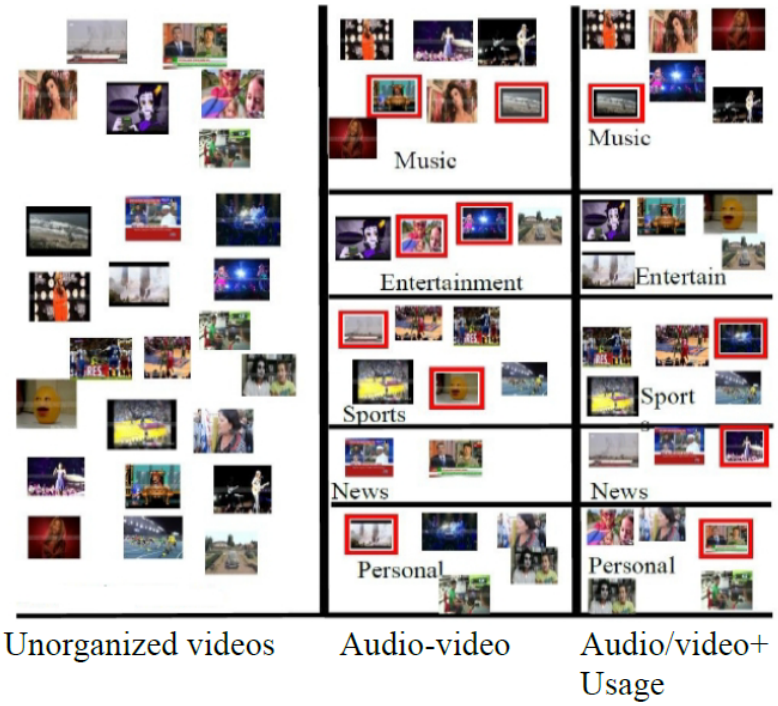

Fig. 6. First column shows the unorganized videos. The second column denotes accuracies obtained with audio-video features alone while third column shows performance after fusing usage with audio-video features. Frames marked in red indicate classification errors.

3. Crane, R., Sornette, D.: Viral, Quality and Junk Videos; Separating Content from Noise in an Information Rich Environment. AAAI Symposium on Social Information, (2008)
4. Filippova, K., Hall, K.: Improved Video Categorization from Text Metadata and User Comments. Annual SIGIR Conference, (2011)
5. Pentland, A.: Social Signal Processing. Signal Processing Magazine (2007)
6. Sargin, M., Aradhye, H.: Boosting Video Classification Using Cross-video Signals. ICASSP(2011)
7. Song, Y., Zhang, Y., Zhang, X.,Cao, J., Li, J.: Google Challenge- Incremental Learning for Web Video Categorization on Robust semantic Feature Space. 17th ACM Conference on Multimedia (2009)
8. Song, Y., Zhao, M., Yagnik, J., Wu, X.: Taxonomic Classification of Web-based Videos. CVPR (2010)
9. Tsoularis, A.: Analysis of logistic growth models. Res.Lett.Inf.Math.Sci, (2001)
10. Wu, X., Zhao, W., Song, Y.,Kumar, S., Li, B.: YouTubecat-Learning to Categorize Wild-web Videos. CVPR (2010)
11. Zhou, X., Huang, T.: Relevance Feedback in Image Retrieval: A Comprehensive Review. Multimedia Systems Special Issue on Content Based Image Retrieval (2003)