Who Produced This Video, Amateur or Professional?

Jinlin Guo, Cathal Gurrin
CLARITY and Computing
School, Dublin City University
Glasnevin, Dublin 9, Dublin,
Ireland
{jinlin.guo, cgurrin}@computing.dcu.ie

Songyang Lao
School of Information System
and Management
NUDT, Changsha, Hunan,
China
laosongyang@vip.sina.com

ABSTRACT

As the increasing affordability for capturing and storing video and the proliferation of Web 2.0 applications, video content is no longer necessarily created and supplied by a limited number of professional producers; any amateur can produce and publish his/her video quickly. Therefore, the amount of both professional-produced as well as amateur-produced video on the web is ever increasing. In this work, we propose a question; whether we can automatically classify an Internet video clip as being either professional-produced or amateur-produced? Hence, we investigate features and classification methods to answer this question. Based on the differences in the production processes of these two video categories, four features including camera motion, structure, audio feature and combined feature are adopted and studied along with with four popular classifiers KNN, SVM GMM and C4.5. Extensive experiments over carefully-constructed, representative datasets, evaluate these features and classifiers under different settings and compare to existing techniques. Experimental results demonstrate that SVMs with multimodal features from multi-sources are more effective at classifying video type. Finally, for answering the proposed question, results also show that automatically classifying a clip as professional-produced video or amateur-produced video can be achieved with good accuracy.

Categories and Subject Descriptors

I.4 [**Image Processing and Computer Vision**]: Miscellaneous; I.5.4 [**Pattern Recognition**]: Computer Vision—*Applications*

General Terms

Experimentation, Performance

Keywords

Amateur-produced Video, Professional-produced Video, Camera Motion, Video genre Classification

.

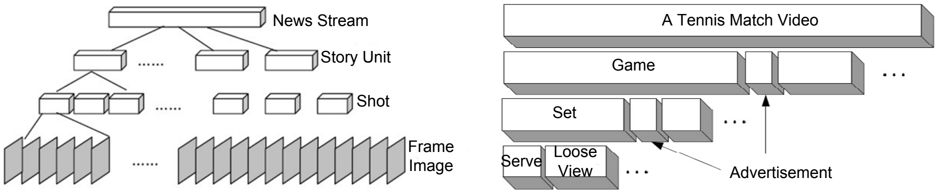
1. INTRODUCTION

Video content has been historically created and supplied by a limited number of production companies, TV networks, and cable networks. They are produced by professionals and consumed by general public. The video quality was in general guaranteed since it was recorded by professional capturing device and subject to careful post-produced according to certain cinematic principles.

Nowadays, the increasing affordability for capturing and storing video has resulted in a massive amount of personal video content, and the proliferation of Web 2.0 applications is re-shaping the video consumption model. Especially, the rise of some easy-to-use social networking websites such as YouTube [1], makes it easy for users uploading, managing, sharing video. Therefore, the users on the Internet are no longer only video consumers, but also participators and producers, just as the slogan of Tudou [2], one of the most popular video sharing websites, "Everyone is the director of life". Now, hundreds of millions of Internet users are self-publishing consumers. This results in an explosive increase in the quantity of Internet video. Recent statistics show that, on the the primary video sharing website YouTube, 48 hours of video are uploaded every minute by users, resulting in nearly 8 years of content uploaded every day, and more video is uploaded to YouTube in one month than the 3 major US networks created in 60 years[3]. The video on the Internet, we call them user-uploaded video, may be either produced by amateur or professional based on its original producer (the user uploaded the video may be the authors of the video or not). Hence, the user-uploaded video can be categorized into amateur-produced video (APV) and professional-produced video (PPV) based on the author type.

We define an APV clip as being recorded by an amateur without much knowledge in producing video and generally using personal video capture devices, then uploaded to websites by the user (it may be the amateur or not) with little post-production. By contrast, the PPV is captured by professional devices and edited based on certain cinematic principles, such as news, sports and movies. Note that, a number of Internet video clips are made by extracting/ripping content from PPV such as TV programs, DVD movies, and then uploaded (sometimes even with some captions and background music added). In this work, we still consider them as PPV. Therefore, compared with PPV, the APV has the

[1] http://www.youtube.com
[2] http://www.tudou.com/
[3] http://www.youtube.com/t/press_statistics



(a) A structural method for broadcasted news video

(b) A structural method for a tennis match video

Figure 1: Two structural methods for broadcasted news video and tennis match video respectively

following characteristics [7].

- A great number of APV clips can be found on the Web. Since APV requires less production efforts, anyone can readily make short video clips by using a camcoder or even a smartphone. Easy of production creates a massive amount of APV content.
- Due to the uncontrolled capturing conditions and accompanied personal capture devices, APV is most of the time of lower quality than PPV.
- The APV is usually less structured: APV is not as well structured as PPV. The PPV clips are consumed by general public. They are produced following certain cinematic principles. The structure is understood such as the structures of news video and tennis video shown in Fig. 1. However, APV is usually captured by different amateurs, who generally do not follow professional guidelines and best practice when producing video. In most cases, there is no post-production before uploading to the video sharing websites. Therefore, there is less structure information existing in the APV.

The question raised here is whether we can automatically determine that a video clip was created by an amateur or professional? That is, can we classify an Internet video clip into PPV and APV automatically? This is helpful, for example, when generating ranked lists in response to a user query or when categorizing results. For example, a user searches for a concert video clip and prefers the clips published by official producer, rather than by the audience in the scene. In this work, we investigate features and techniques for answering this question. Our approach is based on the differences that are inherent in the production processes of these two video categories. Multimodal features including camera motion, structure information, audio feature and combined feature together with four popular classifiers are adopted and also evaluated within different experiment settings. Furthermore, with the goal of comparitive evaluation, two representative state-of-the-art frameworks are also implemented to classify APV and PPV.

The paper is structured as follows. In section 2, we briefly review the related work. In section 3, we detail the multimodal features and algorithms for classifying a user-uploaded video clip into PPV or APV. The experiment and results are presented and summarized in section 4. Finally, Section 5 concludes this work and outlines the future work.

2. RELATED WORK

Most relevant work in our context focuses on video genre classification. Some researchers have focused on classifying segments of video, such as identifying violent [10] or scary [18] scenes in a movie. However, most of the video classification work attempts to classify an entire video clip into one of several genres, such as *sports*, *news*, *cartoon*, *music*. In general, the previous methods can be categorized into four types: text-based approaches [1, 23], audio feature based approaches [5, 15, 17], visual feature based approaches [4, 20, 22], and those that used some combination of text, audio and visual features[3, 4, 8]. In fact, most authors incorporated audio and visual features into their approaches (we call it content-based approaches), and these approaches achieved good performance. Here we will give a brief review about the content-based approaches. Extensive surveys of these techniques can be referred in [2, 12].

The combination of audio and visual low-level features attempts to incorporate the audio and visual aspects that these features represent and complement each other. Audio features can be derived from either the time domain or the frequency domain. Time domain features such as the root mean square of signal energy (RMS), Zero-Crossing Rate (ZCR) and frequency domain feature MFCC are commonly used in previous work [3, 12].

Visual features in general include motion features [11], keyframe image features such as Scale Invariant Feature Transform (SIFT) [22] and color or texture [3, 4, 8], structure features, such as average shot length, gradual transition and cut shot ratios [8, 12, 21], identification of some simple objects [20], with research focussed on how to combine these features.

Ways of using these features investigated in existed work include many of the standard classifiers because of their ubiquitous nature, such as KNN [8, 22], Linear Discriminant Analysis (LDA) [8], SVM [3, 4, 5, 8, 16, 21], C4.5 decision tree [4, 9], GMM [11, 12, 17, 20]. Moreover, some more complicated methods such as HMM [4, 19, 20] and neural networks [12] were also introduced to video genre classification.

It should be noted that two evaluations strongly promoted the research on Internet video genre classification. The first one is set out by Google as an *ACM Multimedia Grand Challenge task* in 2009 (also in 2010) [4]. Followed that was the *Genre Tagging Task* in MediaEval 2011 [5], which focused on genre classification of Internet video.

The Internet video can be categorized into a number of genres. For instance, the genre classification defined in Google Video consists of 38 genres, such as *business*, *music*, *news*, *sports* and so on [6]. However, in this work, we focus on determining if a video clip is APV or PPV, namely if it is

[4] http://www.sigmm.org/archive/MM/mm09/MMGC.aspx.htm

[5] http://www.multimediaeval.org/mediaeval2011/

[6] http://video.google.com/genre.html

produced by amateur or professional.

Our approaches are based on the features including background or camera motion, structure, audio and combined feature that can discriminate the APV and PPV. Compared with previous work, our contributions in this work are that 1) we propose the question of recognizing the APV and PPV; 2) four commonly used classifiers with four features especially background or camera motion are evaluated with extensive experiments over carefully selected datasets, moreover, we compare and evaluate our approaches with two representative techniques in previous work to address this problem.

3. CLASSIFYING AMATEUR AND PROFESSIONAL PRODUCED VIDEO

In this section, we will describe the features and classifiers used for classifying APV and PPV in this work. Firstly, in section 3.1 the background or camera motion features are described. Structure information and audio features are described in section 3.2 and section 3.3 respectively. In section 3.4, the classification algorithms adopted are introduced.

3.1 Camera Motion Feature

As stated, the APV is generally recorded by person without much knowledge about cinematic principles using personal devices under uncontrolled capturing conditions, then uploaded to the sharing websites with less post-production (such as stabilization). Therefore, APV is apt to suffer from more irregular camera motion than PPV. The camera motion feature is likely to be a good potential discriminator between APV and PPV.

The visual quality of video is highly relevant to three properties of camera motion (CM) [6], that is, *speed*, *direction* and *acceleration*. These properties affect video quality in different ways. If the speed of CM is high, the captured frames will be blurred. When the speed is normal, but the direction of CM changes frequently, namely, the camera moves back and forth repeatedly, the captured video is regarded as shaky. When speed is normal and direction is consistent, but the accelerations of CM in consecutive frames are uneven, that is, the variance of acceleration is large, the captured video is inconsistent. The normal CM with few direction changes and steady accelerations lead to stable video.

Specifically, we adopt the block-match based optical flow approach in [6] to detect the background or camera motion features, since it is computational efficient. For a video clip c, by the method in [6], a set of motion vectors is obtained, set as $\boldsymbol{V} = \{\boldsymbol{v}_{k-1,k}\}_n$, where, $\boldsymbol{v}_{k-1,k}$ is the CM vector extracted from the consecutive frame $k-1$ and k, n is the number of motion vectors extracted in c. Based on $\boldsymbol{V}$, a set of acceleration vectors can be calculated, set as $\boldsymbol{A} = \{\boldsymbol{a}_{k-1,k,k+1}\}_{n-1}$, here,

$$\boldsymbol{a}_{k-1,k,k+1} = (\boldsymbol{v}_{k-1,k} - \boldsymbol{v}_{k,k+1})/\Delta t \doteq \boldsymbol{v}_{k-1,k} - \boldsymbol{v}_{k,k+1} \quad (1)$$

where Δt is time interval between two consecutive-extracted frames. Since we sample the frames uniformly (five frames per second), that is, Δt is a constant. Meanwhile, a set of direction changes is obtained, set as $\boldsymbol{\theta} = \{\theta_{k-1,k,k+1}\}_{n-1}$ where $\theta_{k-1,k,k+1}$ is the direction change, namely the angle between $\boldsymbol{v}_{k-1,k}$ and $\boldsymbol{v}_{k,k+1}$, calculated by:

$$\theta_{k-1,k,k+1} = arccos\left(\frac{\boldsymbol{v}_{k-1,k} \cdot \boldsymbol{v}_{k,k+1}}{\|\boldsymbol{v}_{k-1,k}\| \, \|\boldsymbol{v}_{k,k+1}\|}\right) \quad (2)$$

We choose the mean, second order central moment (or variance), third order central moment and fourth order central moment of $\boldsymbol{V}$, $\boldsymbol{A}$ and $\boldsymbol{\theta}$ as the camera motion feature since these statistics represent the change and distribution properties of CM in clip c. Specifically, for $\boldsymbol{V}$, the mean is computed as:

$$\bar{\boldsymbol{v}} = \frac{\sum \boldsymbol{v_{i-1,i}}}{n} \quad (3)$$

The t_{th} order central moment is calculated as:

$$\boldsymbol{m}_t = \frac{\sum (\boldsymbol{v}_{i-1,i} - \bar{\boldsymbol{v}})^t}{n} \quad (4)$$

where $t \in \{2, 3, 4\}$.

For $\boldsymbol{A}$ and $\boldsymbol{\theta}$, the same statistics are calculated, but with the number of $n-1$. Finally, a 20 (by concatenating the mean, second order central moment, third order central moment and fourth order central moment of CM vector, acceleration vector and direction change value, namely, $4*2+4*2+4$) dimensional feature vector is obtained to represent the video clip c.

3.2 Structure Information

As stated in the Introduction, APV is not usually as well structured as PPV. Structure or temporal information is strongly related to PPV genre, e.g. *business* and *music* clips tend to have a high visual tempo, *business* uses a lot of gradual transitions etc. Therefore, structure information from the shot may help discriminate APV and PPV.

We extract structure information including *shot number*, *average shot length*, *cut shot ratio*. Here, we apply the shot boundary detector in [14], which finds two types of shot boundaries,i.e. cut and gradual transition. The average shot length is computed by averaging all the shot lengths in a video. Additionally, we calculate the ratio of cut shot to the overall shot boundaries.

3.3 Audio Feature

Audio information has the potential to be an important cue discriminating different video genres. Most of the common video genres have very specific audio signatures, e.g. in news there are a lot of monologues/dialogues, sports have a mixture of commentator speaking, applause and clapping, and movie contains a mixture of soundtrack and dialogues, etc.. However, because of the open and sharing of the video websites and the diversity of amateurs, APV can be about anything in any scene, and anyone can be a star, from lip-synching amateurs to skateboarding dogs. Therefore, the accompanied audio content in APV may be more complex and diversified. Audio features such as RMS, ZCR and MFCC are commonly used for video genre classification in previous work, especially the MFCC. In previous work, MFCC features directly or their statistics such as mean and standard deviation were used for video classification [3, 15]. In this work, we will use MFCC together with the bag-of-audio-word (BoAW) representation following the method in [5]. The BoAW is derived from the popular bag of word in text-document classification.

The process is as follows. Firstly, the signal is sampled at 16kHz, then MFCC features are calculated over 25ms windows/frames every 10ms. The "null" MFCC, which is proportional to the total energy in the frame, is also included. Furthermore, delta coefficients and acceleration coefficients, which estimate the first and second order derivation of MFCCs respectively and exhibits the dynamic characteristic of the audio content, are also adopted. In total, the extraction of MFCCs results in a 39-dimensional feature vector for each frame. Then, each video's accompanied audio is represented as a set of $d = 39$ dimensional MFCC feature vectors, where the total number of frames from an entire video depends on its duration.

In order to create the BoAW representation, a vocabulary with 2,000 audio words is created by K-means clustering on a randomly sample 500,000 MFCC feature vectors. Finally, all features of a video's soundtrack are assigned to their closest (using Euclidean distance) audio words. This produces histograms of audio word occurrences for each video clip, and are then used as feature input for classifying the APV and PPV.

3.4 Classification Algorithms

Some complicated methods such as HMM and neural networks were employed in previous work. However, they need much more time and computational effort to train classification models. In the context of this work, determining a video clip as APV or PPV is typically a binary classification question. Therefore, in this work, four popular and relative easy-to-perform classification approaches are selected for our evaluation experiments, namely, K-Nearest Neighbors (KNN), Support Vector Machines (SVM), Gaussian Mixture Model (GMM), and C4.5 decision tree.

KNN classifier generates clusters representing the classes of feature points and assigns a feature instance to the cluster that has k instances closest to it. In our work, cosine distance is adopted for KNN method and k is set as 1. SVMs map an input space into a high dimensional feature space through a kernel function and then constructs the optimal separating hyperplane in the high dimensional feature space. With respect to the SVM kernel, the Gaussian Radius Basis Function (RBF) kernel is used since it is widely used and always achieves good performance across different applications. When using the GMM method, one model is trained for one class of video. The Expectation-Maximization (EM) algorithm is adopted to estimate the parameters of a GMM. When testing, a sample is predicted to the class whose model outputs larger confidence. In this work, 9 mixture components are used in GMM. The C4.5 decision tree recursively subdivides a set of data by using the concept of entropy from information theory. The feature which provides the most information gain, as defined by the difference in entropy, at each recursion is used to form a decision based on the values of the feature. The result is a tree where each node has a feature and a decision depending on its value.

4. EXPERIMENTS AND RESULTS

4.1 Experimental Dataset

To evaluate the efficiency of the features and methods used in the work, we select about 150 hours', 2,000 video clips, each of which either belongs to APV or PPV. The duration of each video clip is less than 10 mins.

In the 2,000 video clips, 500 are annotated as APV, while the others are labeled as PPV. The APV clips are from the NIST TRECVID 2011 Multimedia Event Detection (MED) task [13]. The MED dataset consists of publicly available video content posted to Internet video hosting sites. The annotation work is carefully conducted. Each video clip is viewed by three annotators. One clip is deemed as APV only when all three annotators consider it as APV. The annotations mostly rely on the semantics of video content, such as the scenes, dialogues in video. The annotations show that the annotators easily obtain consistency in most cases. The PPV clips were crawled from YouTube, including three video genres, news, sports, and movies. We checked the collected PPV clips carefully. Finally, 1,500 clips were selected, 500 clips for each genre. We determine a video as news or sports video only based on the TV channel logo, such as RTE NEWS, BBC, CCTV-4. The movie clips are fragments from multiple movies.

The experiments focus on the performance among features, video genres and classification methods. Besides the three single features namely camera motion feature (CMF), structure feature (SF) and audio feature (AF), we also evaluate the combined feature (CF) that are from the concatenation of the three single features and are normalized before input as features for classifiers. Firstly, we investigate the performance when only considering to classify a specific PPV genre and APV. Then experiment for discriminating any genre of PPV clip and an APV clip is implemented. Hence, we identify four experiments:

- News *vs.* APV.
- Sports *vs.* APV.
- Movies *vs.* APV.
- Mixture *vs.* APV.

The following experiments are conducted using five-fold cross validation. The mean accuracy and standard deviation over five-fold cross-validation are reported.

4.2 News *vs.* APV

The first experiment evaluates the performance of the four features with different classification methods when assuming the PPV is only news video. Experiments are conducted over the datasets including the 500 news clips and the 500 APV clips. In Table 1, we show the results. It should be noted that, the mean accuracies reported by SVM classifiers are based on the best parameters trained on the entire 1,000 video clips using five-fold cross validation.

Table 1 illustrates that in the three single features, CMF reports the best performance for all four methods. Especially compared with the structural feature (SF), the CMF achieves much better performance. Moreover, CMF also attains comparable accuracies with the combined feature (CF). When using GMM and C4.5, the CMF even outperforms the CF by a minor value. In the two cases where CMF is outperformed by CF when using KNN and SVM classifiers, the performance difference is marginal. The excellence of CMF for discriminating the news video and the PPV is in line with the differences between the two video categories. News video are usually recorded by experienced photographers. When capturing a news clip, the camera is kept moving uniformly in most cases, that is, the camera moves toward a certain direction with relative uniform speed. However, camera motions in APV clips are more

Table 1: Accuracy(%) comparison on four features with four classification methods when the PPV only contains news video

	KNN	SVM	GMM	C4.5
CMF	91.4 ± 3.13	92.0 ± 2.33	88.3 ± 2.23	89.5 ± 2.21
SF	78.8 ± 1.68	82.3 ± 3.02	81.2 ± 2.91	85.8 ± 2.89
AF	90.3 ± 2.89	91.3 ± 1.92	82.3 ± 2.71	78.3 ± 1.99
CF	94.3 ± 2.20	92.4 ± 2.22	87.7 ± 2.20	88.4 ± 2.11

irregular since the APV clips are typically captured by personal easy-shaking small-size devices such as smartphones. Furthermore, post-processings such as stabilization may be performed on news video before broadcasting [7].

Audio feature (AF) also attains high accuracies with KNN and SVM algorithms. This may be attributed the audio content accompanied in news video is mainly from persons, such as the anchorpersons, interviewees or dialogs. In contrast, the audio content in APV can be anything, such as person voice, dog barking, music and so on. AF reports much reduced performance when using GMM and C4.5 methods compared with KNN and SVM. We consider that this because the GMM and C4.5 learns less discrimination information when using much higher dimensional audio features with BoAV representation (2,000 dimensions). Furthermore, KNN and SVM classifiers attain best performance overall.

4.3 Sports *vs.* APV

As a second experiment, we focus on discriminating the sports video and the APV. Results are shown in Table 2. It is worth noting that the sports video clips we collected from YouTube are mainly soccer, rugby and tennis matches. This table shows that the adopted features are less powerful for discriminating sports and APV than discriminating news and APV. Moreover, the table also shows a clear advantage of using audio feature (AF) over the other two single features. For all the classifiers, AF attains the best performance of the three single features. We speculate that in the sports video (especially in soccer, rugby and tennis matches), accompanied audio categories mainly include commentator speaking, applause and clapping from the audiences in live. However, the audio in APV is more diversified.

CMF report accuracies around 80 for all classifiers. The best is 82.1 ± 1.73 with GMM, whereas 78.3 ± 2.20 with KNN is the worst. Compared with APV, camera motions are also very strong in sports video, especially in field sports such as soccer matches. Several camera motion modes such as *zoom in/out*, *pan* and *tilt* are commonly existed in sports video, which makes the CMF is less discriminative for sports and APV.

Structure information is not sufficient to distinguish the sports video and APV. We attribute this most to the fact that complex camera motions in sports video result in poor performance in shot boundary detection. Nevertheless, the combined feature (CF) achieves the best performance nearly for all classifiers, except when using GMM classifier. This proves that the three features are complementary and boost the classification accuracies.

Table 2: Accuracy(%) comparison on four features with four classification methods when the PPV only contains sports video

	KNN	SVM	GMM	C4.5
CMF	78.3 ± 2.20	81.3 ± 1.63	82.1 ± 1.73	78.7 ± 2.63
SF	71.8 ± 2.18	72.6 ± 2.19	68.6 ± 2.67	77.2 ± 2.54
AF	87.4 ± 1.79	85.7 ± 2.31	83.3 ± 2.02	78.9 ± 2.27
CF	89.4 ± 2.49	87.2 ± 1.82	82.8 ± 1.70	83.2 ± 2.73

Table 3: Accuracy(%) comparison on four features with four classification methods when the PPV is only movie video

	KNN	SVM	GMM	C4.5
CMF	70.1 ± 3.73	78.1 ± 4.11	77.2 ± 2.73	76.1 ± 3.84
SF	77.6 ± 3.18	75.9 ± 2.89	78.3 ± 3.27	74.1 ± 4.54
SF	71.2 ± 2.92	76.2 ± 3.31	73.2 ± 3.02	73.6 ± 3.27
CF	81.5 ± 3.19	85.8 ± 3.34	82.6 ± 2.91	82.3 ± 2.96

4.4 Movies *vs.* APV

In an attempt to evaluate the power of the four features as well as four classifiers for discriminating movies and APV, we conduct the third experiment, in which, the PPV clips are all fragments from films. The movies collected for this experiments include clips from romance films, action films, horror films and thrillers.

The performance of the PPV only movies is listed in Table 3. Results show that the three single features report close performance with the four classification algorithms. For CMF, the best accuracy is 78.1 ± 4.11 with SVM classifier. With respect to structure information, accuracy attains the best of 78.3 ± 3.27 when using GMM method. Whereas, top accuracy is 76.2 ± 3.31 when adopting AF with SVM. Overall, the accuracies are around 75 but less than 80 when using single features. Another observation is that classifying movies and APV results in much more significant standard deviations. The intra-class differences in these selected movies are apparent. We only take audio features as example, in romance films, music and dialogs may be the main audio types. However, in horror films, synthetic horror sound is one of the most important "actors". Therefore, huge intra-class differences in these selected films results in significant performance variance.

Furthermore, performance of structure feature (SF) proves competitive to the other two single features. When using KNN and GMM methods, the SF yields higher performance than CMF and AF. We credit this to the structure information being strongly related to movie genre, e.g. action film clips tend to have a high visual tempo, romance movies use a lot of gradual transitions,etc.

Finally, when combining these simple features into composite features, accuracies are boosted significantly. Again, we deem this performance gain comes at the enhancement of discriminative power of the combined features (CF).

4.5 Mixture *vs.* APV

With the goal of answering the question as stated in the Introduction, the fourth experiment is performed under the

[7] For YouTube, the users are required to stabilize the video using the stabilizer tool in the YouTube Video Editor before uploaded it. However, firstly, the effectiveness of this tool on APV needs further evaluation. Secondly, the APV clips used in this work are not necessarily from YouTube

Table 4: Accuracy(%) comparison on four features with four classification methods when the PPV is mixture of multiple PPV genres

	KNN	SVM	GMM	C4.5
CMF	72.2 ± 3.03	80.7 ± 2.97	73.7 ± 3.13	72.4 ± 3.72
SF	75.1 ± 3.27	77.2 ± 3.39	78.6 ± 3.25	69.8 ± 3.64
AF	77.4 ± 2.89	78.7 ± 3.42	75.1 ± 2.92	70.9 ± 2.73
CF	82.6 ± 3.17	85.3 ± 2.88	80.8 ± 2.79	79.1 ± 3.23

condition of the PPV clips are a mixture of clips from different PPV genres. In real scenario, there are much more PPV genres, such as those 38 genres defined in Google Video. In this work, we consider that the PPV consists of uniform mixtures of news, sports and movies. However, considering more PPV genres and random mixtures of PPV clips can be inspired by the experiment here and would be an obvious future research task.

In order to perform these experiments on balanced datasets, we divide 1,500 PPV clips into three uniform parts randomly. In practice, 500 clips for each PPV genre are divided into three uniform parts (167+167+166) randomly. Then three groups of 500 video clips are obtained by mixing three parts from each genre respectively. Experiments are performed on each group of 500 PPV clips versus the 500 APV clips with cross validation. Means accuracy and standard deviation are calculated over the performance from the cross validations.

Results are shown in Table 4. Similar observations as in section 4.4 can be found. When using the single features, the performance differences are marginal and most of the accuracies are between 70 to 80. This may be attributed the fact that none of these single features are sufficient to discriminate each PPV genre and APV. Because of the diversity in the collected PPV clips, significant standard deviations are also reported. Moreover, when combining these simple features into composite features, accuracies are boosted significantly.

When comparing four adopted classification algorithms, we find SVM classifiers outperform other methods when using three of four features. The lower two accuracies are reported by C4.5 decision trees with structure feature (SF) and audio feature (AF), namely 69.8 ± 3.64 and 70.9 ± 2.73 respectively, which are much worse than those obtained by the other three methods when using the same features. However, when using the combined feature (CF) as input for C4.5, the performance is significantly improved.

4.6 Summary of Experimental Results

In order to compare performance of the features and classification algorithms adopted, we aggregate the accuracies of the four features and four algorithms in the four experiments above, as shown in Table 1∼4 . Table 5 compares the discrimination of these four features, in which, each value is the average of the four mean accuracies of corresponding feature in relevant experiments. Table 6 compares the performance of these four classifiers, in which, each value is the average of the four mean accuracies of corresponding classifier in relevant experiments.

Table 5: Averages of the evaluated four features on discriminating different PPV genres with APV.

	News vs.	Sports vs.	Movie vs.	Mixture vs.
CMF	90.3	80.1	75.4	74.5
SF	82.0	72.6	76.5	75.2
AF	85.6	83.8	73.6	75.5
CF	90.7	85.7	83.1	82.0

Table 6: Averages of the evaluated four classification algorithms on discriminating different PPV genres with APV.

	News vs.	Sports vs.	Movie vs.	Composite vs.
KNN	88.7	81.7	75.1	76.8
SVM	89.5	81.7	79.0	80.5
GMM	84.9	79.2	77.9	77.1
C4.5	85.5	79.5	76.5	73.1

Our first observation is that none of these single features are sufficient to discriminate each specific PPV genre and APV. When comparing these single features, we find CMF is better for distinguishing the APV with the PPV genres whose structure is clear and camera motion is simple and regular, such as news video. AF may yield better performance when it is used for discriminating APV with PPV genres which contain less audio types, such as news and sports video. Whereas, temporal SF may improves the performance when classifying APV with PPV genres in which, temporal structure is related with tempo or effect, such as movies. On the whole, CMF and AF prove to be more efficient at discriminating PPV and APV, leading to better classification accuracies.

A notable observation we can make is that combination of multimodal features boosts the classification performance significantly nearly for all algorithms in the four experiments. This means that information from multi-sources should be considered when classifying PPV and APV, especially when PPV contains more video genres. This is also in line with the conclusions in previous work [3, 8]. Therefore, adoption of multimodal features appears to be the road map for future work in discriminating PPV and APV.

With respect to specific PPV genres, news video is easier to be distinguished with APV, which may be attributed to the fact that many similar cinematic principles are complied with by different news video producers when producing news video, such as they are generally structure-clear. Therefore, even temporal structure feature yields good performance when classifying news video and APV.

The four classification algorithms show different performance when used with different features and for classifying different PPV genres and APV. KNN and SVM yield the best performance in discriminating news and APV, using camera motion and audio features. They are also good at distinguishing sports and APV when using audio features. Whereas, when using single features, GMM also yields good performance in classifying movie and APV. Another point worth noting is that for each algorithm, the overall performance decreases as the PPV contains more sub-genres of PPV. We illustrate this by Fig. 2

Globally, SVM and KNN yield excellent performance in each experiment with different features. GMM has the advantage of training efficiently. The number of Gaussian components may affect the performance, and for different features, the number may be different. However, in our previ-

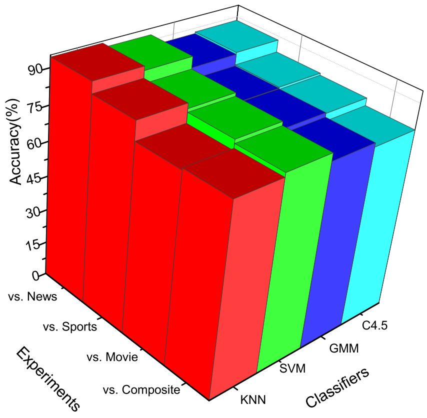


Figure 2: Accuracies of combined features with four classification algorithms in four experiments. It shows that the performance decreases as PPV contains more video genres.

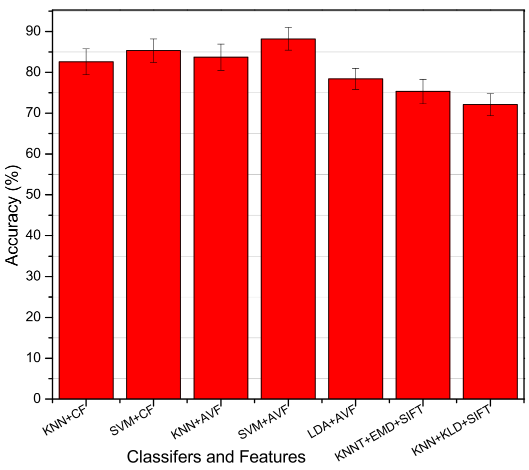


Figure 3: Performance comparison

ous experiments, we set the number as a fixed value 9 for all features and all experiments, which may hinder its performance. The performance of C4.5 relies on the information contained in some discriminative features. Furthermore, when using high dimensional features, such as BoAW representation for audio features, poor accuracies are reported and training is time consuming.

4.7 Extended Experiments

In this section, we apply two state-of-the-art frameworks to the classification of APV and PPV in our context, and also compare our methods with them. One is the framework using audio-visual features (AVF) in [8], the other is the framework with SIFT features in [22]. These two frameworks are representative and originally used for common video genre classification. In [8] tens of audio-visual features, including audio, temporal structure, color, and contour, are combined to classify different video genres such as *news*, *sports*, *music*, *movies* and so on. Three classifiers were used, namely SVM, KNN and LDA. More, recently, there is a trend of using local image features which are scale-, affine- or other-property invariant in the retrieval of imagery and video data, and in [22], Zhang et al. introduced the most popular SIFT feature for video genre classification. With the goals of evaluating these two frameworks in our classification task of APV and PPV, and comparing them with our methods, we perform this extended experiment over the 2,000 clips collected in this work.

Experiments are performed in the same way as in section 4.5. Here, related parameter settings are the same as in [8, 22]. In [8], results from SVM, KNN and LDA with AVF are reported (i.e. SVM+AVF, KNN+AVF, LDA+AVF). In [22], results from KNN with earth mover's distance (EMD) and KullbackLeibler divergence (KLD), histogram representation of 1600-size codebook are also reported (i.e. KNN+ EMD+SIFT, KNN+KLD+SIFT).

Experimental results are shown in Fig. 3. Here, we also list our best performance from SVM and KNN, with the combined features in section 4.5 (i.e. SVM+CF, KNN+CF). It can be seen that, with the task of classifying APV and PPV, combination of multimodal features yields better performance. SVM with audio-visual feature (SVM+AVF) in [8] produces the best accuracy. We deem that the audio-visual feature captures more discriminative information. However, the combined feature (CF) in this work also achieves competitive performance. This may be attributed to the camera motion feature we adopted in this work. In [8], an action feature is also used, however, it is only calculated by assuming that a long shot represents high-action content, whereas short shots mean low-action content. In [8], the audio-visual feature reports much higher accuracies for classifying video genre than the accuracies here, since in that work, the experiments are performed as *one-vs.-all* style. In general, discriminating one specific video genre from a mixture (all) of multi-genres video is easier than classifying two mixtures. This can be validated in our previous experiments that, experiment 1, 2, 3, produce better performance than experiment 4 when using combined features.

From Fig. 3, we can also see that the popularly used SIFT feature reports much worse performance in classifying APV and PPV. In the original work [22], these methods achieve high accuracies, even 100. However, similar to [8], their experiments are also conducted to classify much more specific video genres, e.g. surveillance video captured by 5 different cameras are treated as 5 individual genres respectively, or one specific sports, such as *boxing*, represents one video genre. In our context, PPV includes multiple different video genres, and APV can be about anything. Furthermore, currently successful approaches to video genre classification seems to rely on the application of domain knowledge existing in video genres. Therefore, in our context, this domain-knowledge independent SIFT feature is not capable of discriminating the PPV and APV well.

5. CONCLUSIONS

A large volume of video content is poised to inundate the Internet. In this work, we propose a question of determining the producer of an Internet video as professional or amateur, namely, classifying a video clip as professional-produced video (PPV) or amateur-produced video (APV). Features and classification approaches are also investigated. Based on the differences between the production process of the two types of video, four features including camera motion feature, structure information, audio feature and combined features are adopted together with four popular classifiers KNN, SVM, GMM and C4.5. Four experiments are firstly performed over a carefully selected datasets including 1,500 PPV from YouTube and 500 APV from TRECVid MED datasets. The first three experiments focus on classifying one specific video genre from APV. The fourth one performs discriminating APV and any genre of PPV clip. Furthermore, we further implement two representative techniques in previous work to the task in this work and compare them with our methods. Experimental results demonstrate that SVMs with multimodal features from multi-sources are more effective at classifying the two types of video. Finally, for answering the proposed question, results also show that automatically classifying an clip as professional-produced video or amateur-produced video can be achieved with good accuracy. The future work will focus on two aspects. Firstly, we will conduct this work using more video genres, rather than only news, sports and movies in this work. Moreover, we will also consider evaluating more features from multi-sources and classification methods over larger scale of datasets.

6. ACKNOWLEDGMENTS

Thanks to the Information Access Disruptions (iAD) project (Norwegian Research Council), Science Foundation Ireland under grant 07/CE/I1147 and the China Scholarship Council for funding.

7. REFERENCES

[1] D. Brezeale and D. J. Cook. Using closed captions and visual features to classify movies by genre. In *Proc. MDM/KDD*, 2006.

[2] D. Brezeale and D. J. Cook. Automatic video classification: A survey of the literature. *IEEE Transactions on Systems, Man, and Cybernetics, Part C*, 38(3):416–430, 2008.

[3] H. K. Ekenel, T. Semela, and R. Stiefelhagen. Content-based video genre classification using multiple cues. In *Proc. AIEMPro*, pages 21–26, 2010.

[4] R. Glasberg, S. Schmiedeke, M. Mocigemba, and T. Sikora. New real-time approaches for video-genre-classification using high-level descriptors and a set of classifiers. In *Proc. ICSC*, pages 120–127, 2008.

[5] J. Guo and C. Gurrin. Short user-generated videos classification using accompanied audio categories. In *Proc. AMVA*, pages 15–20, 2012.

[6] J. Guo, C. Gurrin, F. Hopfgartner, Z. Zhang, and S. Lao. Quality assessment of user-generated video using camera motion. In *Proc. MMM*, pages 479–489, 2013.

[7] J. Guo, D. Scott, F. Hopfgartner, and C. Gurrin. Detecting complex events in user-generated video using concept classifiers. In *Proc. CBMI*, pages 1–6, 2012.

[8] B. Ionescu, K. Seyerlehner, C. Rasche, C. Vertan, and P. Lambert. Content-based video description for automatic video genre categorization. In *Proc. MMM*, pages 51–62, 2012.

[9] L. Li, N. Zhang, L.-Y. Duan, Q. Huang, J. Du, and L. Guan. Automatic sports genre categorization and view-type classification over large-scale dataset. In *Proc. ACM MM*, pages 653–656, 2009.

[10] J. Lin, Y. Sun, and W. Wang. Violence detection in movies with auditory and visual cues. In *Proc. CIS*, pages 561 –565, 2010.

[11] P. Martin-Granel, M. Roach, and J. S. Mason. Camera motion extraction using correlation for motion-based video classification. In *Proc. IWVF*, pages 552–562, 2001.

[12] M. Montagnuolo and A. Messina. Parallel neural networks for multimodal video genre classification. *Multimedia Tools Appl.*, 41(1):125–159, 2009.

[13] P. Over, G. Awad, J. Fiscus, B. Antonishek, M. Michel, A. F. Smeaton, W. Kraaij, G. Quénot, et al. TRECVID 2011-An overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proc. NIST TRECVID Worshop*, 2011.

[14] M. J. Pickering, D. Heesch, R. O'Callaghan, S. Rüger, and D. Bull. Video retrieval using global features in keyframes. In *Proc. TREC Video Track*, 2002.

[15] M. Roach and J. Mason. Classification of video genre using audio. In *Proc. Eurospeech*, pages 2693–2696, 2001.

[16] M. Rouvier, G. Linarès, and D. Matrouf. Robust audio-based classification of video genre. In *Proc. INTERSPEECH*, pages 1159–1162, 2009.

[17] M. Rouvier, G. Linares, and D. Matrouf. On-the-fly video genre classification by combination of audio features. In *Proc. ICASSP*, pages 45–48, 2010.

[18] J. Wang, B. Li, W. Hu, and O. Wu. Horror movie scene recognition based on emotional perception. In *Proc. ICIP*, pages 1489–1492, 2010.

[19] J. Wang, C. Xu, and E. Chng. Automatic sports video genre classification using Pseudo-2D-HMM. In *Proc. ICPR*, pages 778–781, 2006.

[20] J. You, G. Liu, and A. Perkis. A semantic framework for video genre classification and event analysis. *Sig. Proc.: Image Communication*, 25(4):287–302, 2010.

[21] X. Yuan, W. Lai, T. Mei, X.-S. Hua, X. Wu, and S. Li. Automatic Video Genre Categorization using Hierarchical SVM. In *Proc. ICIP*, pages 2905–2908, 2006.

[22] N. Zhang and L. Guan. An efficient framework on large-scale video genre classification. In *Proc. MMSP*, pages 481–486, Saint Malo, France, 2010.

[23] W. Zhu, C. Toklu, and S.-P. Liou. Automatic news video segmentation and categorization based on closed-captioned text. In *Proc. ICME*, 2001.