Short User-generated Videos Classification Using Accompanied Audio Categories

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

ABSTRACT

This paper investigates the classification of short user-genera ted videos (UGVs) using the accompanied audio data since short UGVs accounts for a great proportion of the Internet UGVs and many short UGVs are accompanied by single-category soundtracks. We define seven types of UGVs corresponding to seven audio categories respectively. We also investigate three modeling approaches for audio feature representation, namely, single Gaussian (1G), Gaussian mixture (GMM) and Bag-of-Audio-Word (BoAW) models. Then using Support Vector Machine (SVM) with three different distance measurements corresponding to three feature representations, classifiers are trained to categorize the UGVs. The accompanying evaluation results show that these approaches are effective for categorizing the short UGVs based on their audio track. Experimental results show that a GMM representation with approximated Bhattacharyya distance (ABD) measurement produces the best performance, and BoAW representation with χ^2 kernel also reports comparable results.

Categories and Subject Descriptors

I.5 [**Pattern Recognition**]: [Miscellaneous]

General Terms

Short User-generated Video Classification

Keywords

User-generated Video, MFCC, Video Classification

1. INTRODUCTION

As the proliferation of Web 2.0 applications, video is poised to inundate the Internet. Recent statistics show that, on the the primary video sharing website,YouTube, 48 hours of video are uploaded every minute by users, and more video is uploaded to YouTube in one month than the three major US networks (ABC, CBS and NBC) have created in 60 years[1]. In order to manage and retrieve such large amounts of data, work has begun on methods of automatic video categorization, which would make it easier for individuals to retrieve video content matching their needs.

User-Generated Video (UGV) accounts for a major proportion of all WWW video content. Here we define :

DEFINITION 1. (User-Generated Video). *Video that is recorded by an amateur without any professional video editing skills. Differing from professional video material, UGV is often filmed using personal video capturing devices such as mobile phones and digital cameras. This type of video can often be found on content sharing website such as YouTube.*

A distinguished characteristic of UGVs is that they are mostly comprised of short video clips. *Wired Magazine* refers to this small-sized content pop culture as "bite-size bits for high-speed munching" [9]. This is also validated in [3] and statistics of the dataset from the NIST TRECVID 2011 Multimedia Event Detection (MED) task [10]. Chen [3] found that 97.9% of video lengths are less than 600 seconds, and 99.1% are shorter than 700 seconds in their crawled dataset from YouTube. The MED dataset consists of publicly available UGVs posted to various Internet video hosting sites, in total 32,061 clips. As shown in Fig. 1, we have found that more than 47% (15,204) video lengths are less than 60 seconds, nearly 66% are less than 100 seconds. This is mainly due to two reasons. Firstly, common users are more likely to capture short video clips using their own camera. Secondly the video sharing websites generally limit the length of video on regular user uploads, such as the limit of 10 minutes in YouTube.

The UGVs are usually processed with less post-production effort (example the addition of simple music backing) before publishing to video sharing websites [6]. Therefore, compared to long UGVs or professionally produced videos, a large number of the short UGVs are accompanied by single-category audio type (e.g. music). In this work, we propose a new categorization framework for short UGVs (no more than 60 seconds in length) using the seven accompanied audio categories, namely, *person_speaking/talking*, *laughing/clapping /cheering*, *music_scene/background*, *outdoor_urban*, *outdoor_rural*, *indoor_noisy*, and *quiet*. More, we also investigate the techniques for classifying the UGVs into different categories corresponding to different audio categories.

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

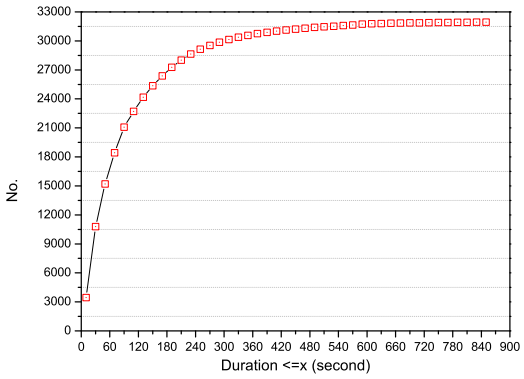

Figure 1: Statistics of duration distribution of UGVs from the TRECVD 2011 MED task dataset

Much research work has attempted to automatically classify an entire video clip into one of several categories, such as *sports*, *news*, *cartoon*, *music*. In general, the previous methods can be categorized into four types: text-based approaches [1, 19], audio feature based approaches [11, 12, 13, 14], visual feature based approaches [5, 16, 18], and those that used some combination of text, audio and visual features[4, 5, 8]. In fact, most authors incorporated audio and visual features into their approaches (we call it content-based approaches); therefore in general, most approaches employ more than one modality. Here we will give a review about the audio feature and content-based approaches, especially the methods using audio features. An extensive review of these techniques can be found in [2].

Roach et al. [12] use the Mel Frequency Cepstral Coefficient (MFCC), from video for genre classification. The authors investigate how many of the coefficients to keep and find that the best results occur with 10-12 coefficients. Classification is performed with a GMM because of its popularity in speaker recognition. The genre studied are *sports*, *cartoons*, *news*, *commercials*, and *music*. In [11], Dinh et al. adopt a Daubechies 4 wavelet to seven subbands of audio clips from TV shows. The features for representing the audio clips utilize the wavelet coefficients from wavelet transformation. Video genre classification was performed using the C4.5 decision tree, kNN, and SVMs with linear kernels. They found that the best performance results were from using the kNN classifier. The genres in this study were *news*, *commercials*, *vocal music shows*, *concerts*, *motor racing sports*, and *cartoons*. Beside low-level audio features MFCC, Perceptual Linear Prediction, M. Rouvier et al. [13] also combine high-level features, namely acoustic instability, speaker interactivity, speech quality to classify the video genres by SVM classifiers. Their experimental results are conducted on a corpus composed from *cartoons*, *movies*, *news*, *commercials* and *music* and they obtain an identification rate of 91%. Furthermore, they further speed up the video genre classification by only considering low-level audio features in [14].

Content-based approaches are found much more often in the video genre classification literature than audio-only approaches, and have attracted more research interest. The combination of audio and visual low-level features attempts to incorporate the audio and visual aspects of the viewing process that these features represent and to complement each other. Visual features in general include motion features, static frame image features, cut (or shot) features, identification of some simple objects, with research focussed on how to combine these features. Some approaches combined all features into a single feature vector while others trained classifiers for each modality and then used another classifier for making the final decision. In [4], beside audio features, visual features including colour and texture descriptors were used. R. Glasberg et al. also used a motion activity descriptor and shot transition descriptor in [5]. Similar features were also used in [8].

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) [2]. Followed that was the *Genre Tagging Task* in MediaEval 2011 [3], which focused on genre categorization of Internet video.

Ways of using these features investigated in existed work include many of the standard classifiers because of the ubiquitousness of them, such as Bayesian, SVM [11, 13], neural networks and GMM [14]. SVM and GMMs methods produced relatively high performance and widely used in previous work.

Accompanied audio based UGV classification is helpful for individuals to retrieve video content when using audio category as filtering. Moreover, it can be also combined to boost content-based video event detection, such as detection of events such as *birthday party* or *wedding* may gain in pre-classified music-related videos. Our contributions in this work are that 1) firstly, we investigate UGV classification using accompanied audio categories; 2) we compare and evaluate three techniques to address this problem. The remaining parts are organized as follows. In section 2, we describe the annotated dataset used in this work, and propose a short UGV classification framework using the accompanied audio categories. Section 3 details our approaches for classifying short UGVs into different categories using audio features. The experimental results are presented and discussed in section 4. Finally, we give our conclusions and outline future work in section 5.

2. DATASET AND UGVS CLASSIFICATION USING AUDIO CATEGORIES

Here, we propose a new classification mode for short UGVs using the audio data, which differs from the genres defined in previous work. We annotated 3,000 videos, in total 26.2 hours, 8.61GB size, which randomly selected from the 15,204 Internet videos that are shorter than 60 seconds in the MED task dataset. To this end, we defined seven audio categories, which are listed and remarked in Table 1, and developed a purpose-built annotation tool for manual annotation. These seven categories are from the definition in previous work [15], also based on our statistics of annotations.

In annotation, each video is divided into multiple 3-second clips, and annotations are conducted at the 3-sec level. The reasons for 3-sec level annotation are that, firstly it assures

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

[3] `http://www.multimediaeval.org/mediaeval2011/`

Table 1: Audio categories defined in this work and the amount of positive samples for each category in the annotated dataset

Categories	Remarks	#
person_speaking/talking (PST)	One or more person are speaking or talking	112
laughing/clapping /cheering (LCC)	They do not necessary occur simultaneously	88
music_scene/background (MSB)	Sound from live or background music	211
outdoor_urban	Sounds in the urban street, from such as car, people walking, talking and so on	97
outdoor_rural	Sounds in the outdoor rural, from such as wind, water, bird tweeting and so on	103
indoor_noisy	Sounds from indoor, from such as when making food.	96
quiet	No audible sound in the video.	107

the audio to be recognizable, secondly, it is less possible that too many different audio categories appear in one clip (in our annotation, most of the clips were annotated 1 or 2 categories), furthermore, it is a compromise between the annotation labor and clip length (actually we can also use these 3-sec level annotation for other research. However in this work, we only use the aggregated entire-video-level annotation). In total, we annotated 32,963 3-sec clips (some are less than 3 seconds since a lot of videos are not integral multiples of 3) divided from 3,000 short UGVs from the MED task dataset. Finally, we aggregate the results on the entire video level. We select the videos accompanied by only one of the defined seven audio categories in this work. Totally, there are 814 videos only accompanied by one audio category (the other UGVs are accompanied at least 2 audio categories). The following experiments are performed on these 814 UGVs, and the number of positive samples for each category is also listed in Table 1.

3. UGVS CLASSIFICATION ALGORITHM USING AUDIO

3.1 Audio Feature Representation

Our fundamental frame-level feature is the MFCCs. MFCC is commonly used in speech recognition and other acoustic classification tasks. For computing the MFCCs, firstly, the signal is sampled at 16kHz, then a short-time Fourier magnitude spectrum is calculated over 25ms windows/frames every 10ms. The spectrum of each window is warped to the Mel frequency scale, and the log of these auditory spectra is decorrelated into MFCCs via a discrete cosine transform (DCT) which approximates the Karhunen-Loeve transform (or equivalently Principal Components Analysis). The "null" MFCC, which is proportional to the total energy in the frame, is also included. Furthermore, the derivation of MFCCs, exhibits the dynamic characteristic of the audio content, which can boost the classification performance. Here, we use the delta coefficients and acceleration coefficients, which estimate the first and second order derivation of MFCCs respectively. 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.

After extracting the MFCC features, the audio of UGVs can be can be represented by sets of MFCC feature descriptors, but the sets vary in cardinality and lack meaningful ordering, which creates difficulties for learning methods (e.g., classifiers) that require feature vectors of fixed dimension as input. To address this problem, a conventional method is encoding MFCC features as a numerical video-level feature vector. We experimented with three different techniques to handle this: 1G, GMM, and BoAW modeling. These fixed-size representations are then compared to one another by several distance measures: the Kullback-Leibler divergence (KLD), approximated Bhattacharyya distance (ABD) and χ^2 distance. The distances between all the video-level feature representations form the input to SVM classifiers.

3.2 SVM Classifiers

The SVM is a supervised learning method used for classification and regression that has many desirable properties [19] and has proven to be a solid choice in several benchmarkings such as TRECVID. Data items are projected into a high-dimensional feature space, and the SVM finds a separating hyperplane in that space that maximizes the margin between sets of positive and negative training examples. Instead of working in the high-dimensional space directly, the SVM requires only the matrix of inner products between all training points in that space, also known as the kernel or gram matrix. Here we use the generalized RBF kernels with different distance measurements, i.e.:

$$K(p, q) = exp(-\gamma \cdot d(p, q)) \quad (1)$$

where $d(p, q)$ is the KLD d_{KLD}, or ABD d_B, or χ^2 distance d_{χ^2} discussed in the following sections.

We use the so-called slack-SVM that allows a tradeoff between imperfect separation of training examples and smoothness of the classification boundary, controlled by a penalty constant c. For both tunable parameters c and γ, we vary the parameters in a grid search area of $[2^{-5}, 2^7] \times [2^{-7}, 2^3]$ and choose the ones that maximize classification accuracy via five-fold cross validation. When testing, the resulting distance-to-boundary is a real value that indicates how strong ly the video is classified to the category. The test videos are then ranked according to this value. Following conventions in information retrieval, we evaluate classifiers by calculating their average precision (AP), since it not only considers precision but also recall by taking into account rank position.

3.3 Single Gaussian Modeling (1G)

The rationale of employing a 1G model is that different types of sounds whose average spectral shape and variation, as calculated by the cepstral feature statistics, should be sufficient to discriminate categories. Specifically, to describe a video's MFCC features as a single feature vector, we ignore the time dimension and treat the set of as a "bag of the frames" in MFCC feature space, which we then model as a single, full-covariance Gaussian distribution. This Gaussian is parameterized by its 39-dimensional mean vector μ and 39×39-dimensional covariance matrix Σ.

To calculate the distance between two Gaussians, as required for the gram-matrix input (or kernel matrix) for the SVM, we adopt the KLD measures. If two videos p and q are modeled by single Gaussians as:

$$p(x) = N(x; \mu_p, \Sigma_p),\ q(x) \sim N(x; \mu_q, \Sigma_q) \tag{2}$$

respectively, then the distance between two videos is taken as the KLD between Gaussians $p(x)$ and $q(x)$ i.e:

$$\begin{aligned} d_{KL}(p,q) = (\mu_p - \mu_q)^T(\Sigma_p^{-1} + \Sigma_q^{-1})(\mu_p - \mu_q) + \\ trace(\Sigma_p^{-1}\Sigma_q + \Sigma_q^{-1}\Sigma_p) - 2d \end{aligned} \tag{3}$$

where $d = 39$.

3.4 Gaussian Mixture Modeling (GMM)

The distribution of audio (especially, speech) features may not be well fitted by a single Gaussian because of diversity, and in previous work, they are typically modeled by GMM. Here, we also experimented with using a mixture of diagonal-covariance Gaussians, estimate via the expectation maximization (EM) algorithm, to describe the frame-feature distribution. Let the the distribution of two clips $p(x)$ and $q(x)$ be represented by two different GMMs:

$$p(x) = \sum_a \omega_a N(x; \mu_a, \Sigma_a),\ q(x) = \sum_b \omega_b N(x; \mu_b, \Sigma_b) \tag{4}$$

where ω_a, μ_a, and Σ_a are the prior weight, mean, and covariance of each Gaussian mixture component used to approximate clip p, and the b-subscripted values are for clip q. We use the shorthand $p_a = N(x; \mu_a, \Sigma_a)$ and $q_b = N(x; \mu_b, \Sigma_b)$ henceforth for simply notation. To compare the distance of two GMMs, we adopted an approximation to the Bhattacharyya distance that was shown to give good performance in tasks requiring the comparison of GMMs [7]. The Bhattacharyya similarity between two distributions p and q is:

$$\begin{aligned} B(p,q) \stackrel{def}{=} \frac{1}{2}\int\sqrt{pq} = \frac{1}{2}\int\sqrt{\Sigma_{ab}\omega_a\omega_b p_a q_b} \geq \\ \sqrt{\Sigma_{ab}\omega_a\omega_b B^2(p_a, p_b)} = \hat{B}^{vb}(p,q) \end{aligned} \tag{5}$$

where $B(p_a, q_b)$ is the Bhattacharyya divergence between a particular pair of single Gaussians, one from each mixture, namely,

$$\begin{aligned} B(\hat{p},\hat{q}) = \frac{1}{4}(\mu_p - \mu_q)^T(\Sigma_p + \Sigma_q)^{-1}(\mu_p - \mu_q) + \\ \frac{1}{2}log|\frac{\Sigma_p + \Sigma_q}{2}| - \frac{1}{4}log|\Sigma_p\Sigma_q| \end{aligned} \tag{6}$$

To preserve the identity property that $\hat{B}(p,q) = \frac{1}{2}$ if and only if $p = q$. This can be enforced by re-normalizing using the geometric mean of $B(p,p)$ and $B(q,q)$:

$$\hat{B}_{norm}(p,q) = \frac{B^{vb}(p,q)}{\sqrt{B^{vb}(p,p)B^{vb}(q,q)}} \tag{7}$$

The normalized estimation was proved to be a better approximation to Bhattacharyya distance. With this approximation, the corresponding Bhattacharyya divergence is defined as: $d_B(p,q) = -2log(\hat{B}_{norm}(p,q))$. More details can be referred in [7].

3.5 Bag-of-Audio-Word(BoAW)

Inspired by the successes of the popular Bag-of-Visual-Words in computer vision application, we adopt a similar process but on audio features, called BoAW representations. The BoAW representation requires creating an audio vocabulary. Here, we randomly sample 500,000 MFCC features from an extra collection (in the rest of the 3,000 UGVs, but not the 814 UGVs used in this work), and cluster them into V clusters (i.e., "words") using K-means clustering algorithm. K-means may trap into local optimum. In order to overcome this defect, we run it 10 times with different initial centers, and select the one with the least variance.

For the representation, all features of a video's soundtrack are assigned to their closet (using Euclidean distance) audio words (AWs). This produces histograms of AW occurrences which are then used for classification.

For classification when using this BoAW feature representations, we use the χ^2 kernel for SVM. The χ^2 kernel has been reported to outperforms the traditional Gaussian RBF kernel in previous work [17]. The χ^2 kernel is:

$$d_{\chi^2}(p,q) = d_{\chi^2}(H^p, H^q) = \frac{1}{2}\sum_i^V \frac{(h_i^p - h_i^q)^2}{h_i^p + h_i^q} \tag{8}$$

where, $H^p = \{h_i^p\}$ and $H^q = \{h_i^q\}$ are the histograms of AW occurrences for clip p and q respectively.

4. EVALUATION RESULTS AND DISCUSSION

We evaluate the three modeling approaches using five-fold cross validation on the 814 videos, in which, each video only contains one of seven audio categories. At each fold, SVM classifiers for each audio category are trained on 50% of the data, and then tested on the remaining 50%, selected at random.

4.1 Results

The results of the 1G with the KLD measures are shown in Fig. 2. The AP of category *quiet* achieves the best, 0.67, which is much higher than that of other categories. The classification performance of *music_scene/background* (*MSB*) achieves the second best. However, the person-speaki ng related category *person_speaking/talking* (*PST*) is also reported comparable AP even it has much less annotated samples in the dataset. This may attribute to the audio feature we used. Each step in the process of creating MFCC features is motivated by perceptual or computational considerations. It represents the speech amplitude spectrum in a compact form, and models the speech sounds better than the music or other sounds. Therefore, MFCC features are more discriminative for person speaking related audio categories. Videos accompanied by location-related audio categories, *outdoor_rural* and *indoor_noisy*, are also classified well by using this method. Overall, the mean average precision (MAP) achieves nearly 0.35 by using IG+KLD method.

Fig. 3 shows the results using GMM with the ABD measure. Performance changes are also reported when adopting different numbers of Gaussian components between 3 to 12. The performance for four categories increases steadily when the number of components enlarges from 3 to 9. When increasing to 12 components, the APs increase a bit or even

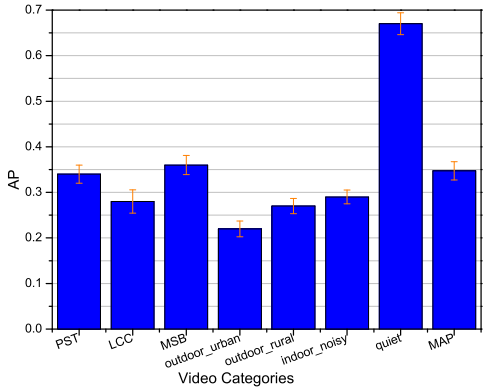

Figure 2: Classification APs using 1G and KLD. Bars and error-bars indicate the mean and standard deviation over five-fold cross-validation testing respectively. MAP is also shown

worsen. On the whole, the MAP of 9-GMM outperforms that of 12-GMM marginally. Therefore, 9-GMM is a good compromise based on the MAP, and ability to discriminate all the audio categories better.

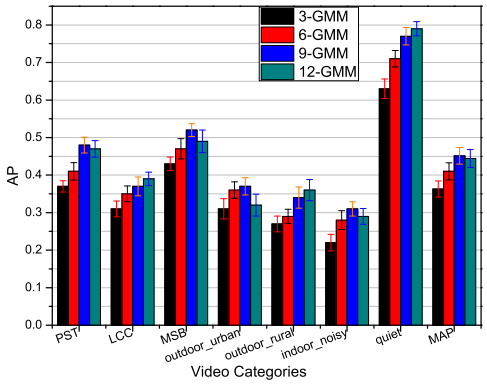

Figure 3: Classification APs using GMM with 3, 6, 9, 12 components and ABD. MAPs are also shown

The performance of BoAW modeling and χ^2 distance measure is shown in Fig. 4. Since the number of AWs plays a important role in the model, we test four various audio vocabulary sizes, namely, 500, 1,000, 2,000 and 4,000. When increasing the vocabulary size from 500 to 2,000, the APs values for four of the seven audio categories increase significantly. Thereafter, the performance gains limitedly or even worsens when increasing the vocabulary size to 4,000. Therefore, audio vocabularies of size 2,000 provides a compromise between computation cost and classification performance.

4.2 Discussion

In Fig. 5, we show the best results from the three modeling approaches. Here we still use the AP as the measurement. Firstly, it should be noted that the classification per-

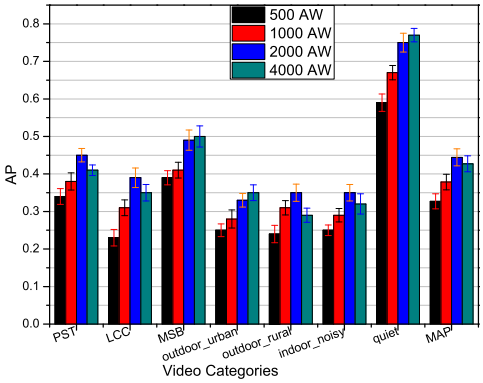

Figure 4: Classification APs using BoAW with 500, 1 000, 2 000, 4 000 AWs and χ_2 distance. MAPs are also shown

formance of category *quiet* reaches much higher than other categories when using any one of the three approaches. Category *quiet* is inherently different from the other categories since no audible sounds in the video, which makes the feature representation of this category be more separable with that of other categories.

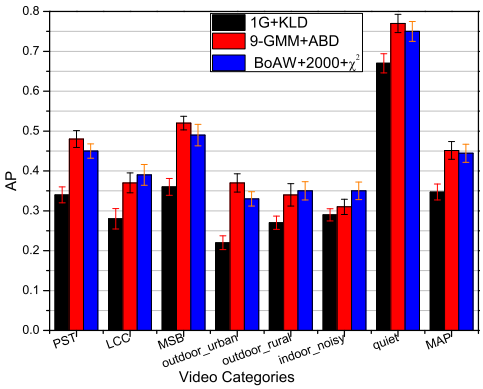

Figure 5: The best APs from three approaches. MAPs are also shown

Our second observation is that the variation of classification performance for different categories. The APs for different categories vary when using the same modeling approach. We deem that the amounts of training data as well as the discrimination of the features affect the performance a lot. Indeed, for category *MSB*, the performance is better than other categories except *quiet* for each modeling method since it has nearly two times of training samples than any of the other categories. However, for category *PST*, it achieves the third best performance, even comparable with category *MSB*, which can be attributed to the more discriminative power of MFCC features for speech audio.

Another observation is that GMM method reports better AP than 1G for all the categories, which means these audio categories are more successful modeled by GMM than by a

single Gaussian. Actually, the audio categories we defined in this work are combinations of multiple subclasses, such as *outdoor_rural* may include sound of wind, water, bird tweeting and so on. This implies that 1G models the specific and more consistent subclasses better, whereas GMM is better for mixed and inconsistent audio categories.

Finally, the GMM method outperforms the method BoAW method, but very marginally. For categories *PST*,*MSB*, *outdoor_urban* and *quiet*, the GMM reports better APs than BoAW, but conversely for the other three audio categories. Overall, the MAP difference reported from these two methods is less than 0.01. This may be explained by the similarity between the BoAW model and the GMM. Such as they usually use an iterative refinement approach to converge to a local optimum, and they both use cluster centers to model the data. The K-means clustering tends to find cluster of comparable spatial extent, while the EM mechanism allows clusters to have different shapes. However, when adopting the appropriate size of BoAW model, it is capable of producing the comparable results as GMM. Therefore, BoAW model may be a efficient option since the iteration in K-means is less complicated.

5. CONCLUSION

In this work, we have investigated the classification of short UGVs that are accompanied by only one single audio category. We define seven types of UGVs corresponding to seven audio categories. We also investigate three modeling approaches for feature representation: 1G, GMM and BoAW. SVM classifiers are trained with three different distance measurements corresponding to three feature representations respectively for kernel learning. Experimental results show that these approaches are effective for categorizing the short UGVs using their accompanied audio. In addition, GMM representation with ABD measurement produces the best performance, and BoAW representation with χ^2 kernel also reports comparable results. Future work will focus on improving the classification performance of short UGVs, testing on larger-scale datasets, and extending to UGVs without duration limitation, finally leading to UGV retrieval based on the accompanied audio categories and multimedia event detection by combining accompanied audio category features.

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. Moreover, many thanks to the Na Li and Yun Jun for their help in the annotation effort.

7. REFERENCES

[1] D. Brezeale and D. J. Cook. Using closed captions and visual features to classify movies by genre. In *MDM/KDD*, San Jose, CA, 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] X. Cheng, C. Dale, and J. Liu. Statistics and social network of youtube videos. In *IWQoS*, pages 229–238, Enskede, Netherlands, 2008.

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

[5] 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 *ICSC*, pages 120–127, Washington, DC, USA, 2008.

[6] J. Guo, D. Scott, F. Hopfgartner, and C. Gurrin. Detecting complex events in user-generated video using concept classifiers. In *CBMI*, pages 177–182, Annecy, France, 2012.

[7] J. R. Hershey and P. A. Olsen. Variational Bhattacharyya divergence for hidden Markov models. In *ICASSP*, pages 4557–4560, Las Vegas, Nevada, USA, 2008.

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

[9] M. Nancy. Manifesto for a new age. *Wired Magazine*, page 128, 2007.

[10] P. Over, G. Awad, M. Michel, J. Fiscus, W. Kraaij, A. F. Smeaton, and G. Quéenot. Trecvid 2011 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2011*. NIST, USA, 2011.

[11] Q. D. Phung, C. Dorai, and S. Venkatesh. Video genre categorization using audio wavelet coefficients. In *Fifth Asian Conference on Computer Vision*, Melbourne, Australia, January 2002.

[12] M. Roach and J. S. D. Mason. Classification of video genre using audio. In *INTERSPEECH*, pages 2693–2696, Aalborg, Denmark, 2001.

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

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

[15] C. C. Tan, Y.-G. Jiang, and C.-W. Ngo. Towards textually describing complex video contents with audio-visual concept classifiers. In *ACM Multimedia*, pages 655–658, Scottsdale, AZ, USA, 2011.

[16] 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.

[17] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, 2007.

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

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