Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 December 2018 doi:10.20944/preprints201812.0086.v1

The User-Pleasant Video Skimming by Multi-Modal Keywords Semantics

Yiqing Shen
Xi'an Jiaotong University, China
syq1220@stu.xjtu.edu.cn

***Abstract*— In this paper, we propose a novel approach of video skimming by exploiting the fusion of video temporal information and keyword information representation extracted from multi-model video information including audio, text and visual indices. In addition, we introduce the brand-safe filtering and sentiment analysis in order to only reserve the user-friendly content in the video skim. In the experiment by using the videos from YouTube-8M dataset, we have proved that the semantic conservation in the video skim from the proposed approach highly outperforms the approaches by only partial information of the video in conserving the semantic content of the video.**

***Index Terms*— Multi-model information fusion, Video skimming, Audio and text classification, keyframe extraction**

I. INTRODUCTION

As the increasing availability of video information, through an immense distribution of sources, people need an automatic system which is capable of analyzing and processing video information. In the last decade, video analytic technology has been developed from these aspects: video shot boundary detection, keyframe extraction[6], motion detection[1] and so on. One of the important tasks is to concise the video to drop redundant information, in which the video skim is a successful research topic representing a brief synopsis of the original video.

Several basic methods of video analysis have been proposed in the last decade. Specific object extraction[4][5], which detects an object from sequential images depending on dynamics features, has been widely used in video information analysis. Keyframe extraction in video summarization using aggregation mechanism[18] or clustering algorithm[19]. What's more, one video summarization approach based on the combination of image and text information has been proposed by M.A. Smith[3]. The authors use keywords extracted from audio signal and scene detection information to generate a video skim. However, it mainly focuses on image information including face detection, text detection, and scene segmentation, which doesn't make full use of audio signal information. Therefore, we try to propose a novel approach from the aspect of conserving as much as possible the semantic information of the video, which fuses all the potential multi-modal video information extracted by multiple deep learning approaches. Furthermore, we propose to filtering out user-unpleasant content which leads to only conserve the user-pleasant semantic content by semantic analysis, because the video skim with unpleasant content corrupts the application of video skims in many cases.

The rest of the paper is organized as follows. In Section II, we propose a novel approach of video skimming, named as Multi-Sources-Indices(MSI). In the approach of MSI, we fuse the video information according keyword occurrence in temporal information with our extraction information from audio, text including speech and visual information by multiple deep learning approaches. In Section III, we present the experimental results by using YouTube-8M [27] dataset with English subtitles. To verify the quality, we calculate the similarity score among the video skimming generated with Multi-Sources-Indices and the original video. We find that the skimming from the proposed approach has significantly outperforms the approaches by only one or two video indices, which means that the proposed approach understand the video information more efficiently in the aspect of semantic meaning. Finally, the conclusion is presented in section IV.

II. MSI VIDEO SKIMMING

In this section we propose a video skimming generation approach, Multi-Sources-Indices (MSI), especially applying to YouTube videos. MSI inventively proposes to convert each information type of video into the time-related keywords information and fuse the information by keyword similarity considering the temporal information in order to extract the video skim by considering both semantic meaning and temporal relation. In order to conform to the current trend of brand safety in the video industry, we exploit Convolutional Neural Network (CNN) [29] text classification and sentiment analysis to filter and ensure the brand safety of the final generated video skim.

We explain the principle to extract the information of each kind of information of video in Section II-A. Text information extraction including speech text is introduced in Section II-C. In Section II-B we exploit the existing algorithm video-MMR[14] to extract video keyframes. Finally, we propose our novel approach of video skimming by fusing the multi-modal information of above and using temporal keywords relations in Section II-D.

A. Audio information extraction

Audio information has a rather influential position in the increasing digital content that is available today. The increasing availability of audio information, through an immense distribution of sources, has led to the need for systems that are capable of automatically analyzing and processing

© 2018 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 December 2018 doi:10.20944/preprints201812.0086.v1

this information. Audio signal processing is an engineering technology that focuses on the computational methods for intentionally altering sounds, methods that are used in many musical applications.

The audio information is a fundamental information especially in the video, for example, the news and the movie. The audio is a significant indication of the important events in the video, even not mentioning that the speech is an important information embedded in the audio channel. We will explain how we segment the into homogeneous regions and classify each segment while use the class name as the important textual keywords in video information fusion.

1) Audio segmentation: Audio segmentation is a very important processing stage for most audio analysis approaches. The goal is to split an uninterrupted audio signal into homogeneous segments. The audio signal is not always meaningfully continuous and might contain segments of silence or noise. Under this inspiration, we prefer audio segmentation in order to identify the useless audio segment like silence and furthermore facilitate the audio classification in order to improve accuracy.

The segmentation part is achieved by removing the silence and detecting the pitch through a semi-supervised approach: a Support Vector Machine (SVM) model [13][15] is trained to distinguish between high-energy and low-energy short-term frames. Toward this end, 10% of the highest energy frames along with the 10% of the lowest ones are used. Then, the SVM is applied (with a probabilistic output) on the whole recording and a dynamic threshold is used to detect the active segments. The principle is to detect the pitch of the audio file using SVM, then according to the preferred parameter, the approach will remove the short segments which is the silence with its pitch lower than the threshold. Although the segmentation step only considers the pitch of the audio, it has effective performance. The result of the audio segmentation is a list of segments, while an example of segmentation is shown in Fig. 1 for the audio example of 20 seconds.[1].

Applying this segmentation model on the audio clip, we get a list of time-points representing pitch change points in time. The audio segmentation is conducted before the audio classification in order to select only the high energy and useful audio segments.

2) Audio classification: Audio classification [16] is the technology classifying the audio segment or shot into predefined classes. In this paper, we would use class names as the semantic text too. To accomplish this goal, we decide to take advantage of the existing Multi-Instance Learning (MIL)[2] model to classify the audio signal.

Multiple Instance Learning is essentially a kind of supervised learning approach. Each audio event contains a set of instances, and each instance presents one feature of the data. In MIL, one event is recognized as a positive event only if it contains at least one positive event. On the contrary, one event is negative when none of the instances is positive. In audio classification, one event presents one audio clip, and each instance presents different feature of the audio signal. We assign an attention parameter to each instance. The class of the event is calculated by the expectation of the class of each instance, which is the product of the probability and the attention parameter. MIL is represented in Eq. 1.

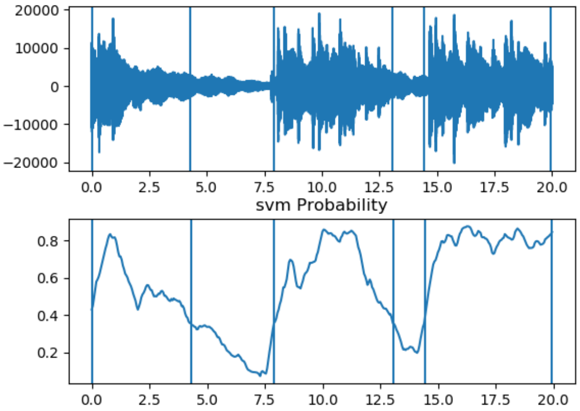

Fig. 1. Scheme for Audio Segmentation

We define a probability space for each bag B_n for each class k. The probability measure a_{nk} satisfies Eq.2, which represents the probability of each instance. The closer a_{nk} is to 1, the more this instance is attended. While the closer a_{nk} is to 0, the less this instance is being considered in the class prediction for this event. For one event, the sum of the attention parameter should be equal to 1.

$$L(B_n) = E_{a_{nk}}(C_k(x)) = \sum_{x \in B_n} C_k(x) a_{nk}(x) \tag{1}$$

$$\sum_{x \in B_n} a_{nk}(x) = 1 \tag{2}$$

where n is the number of the events, k is the number of instances in the event, a_{nk} decides how important this instance is, and C_k presents the classification result of the instance. E represents the function of expectation calculating.

B. Visual information extraction

Keyframe extraction [17] is an effective way to summarize the video. Keyframe refers to the image frame in the video sequence which is representative and able to reflect the summary of a video content. In this paper, we use a keyframes extraction algorithm named Video Maximal-Marginal Relevance(Video-MMR) [14]. Video-MMR iteratively selects keyframes to construct a summary by selecting a keyframe whose visual content is similar to the content of the videos, but at the same time, it is different from the frames already selected in the summary. The selected keyframe is defined as Eq. 3:

[1] https://drive.google.com/open?id=0B2hLTOikyKLJWjBYZ3E2ZEh3 R2JRSUpib254R3ZWdDJPQUZN

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 December 2018 doi:10.20944/preprints201812.0086.v1

$$\begin{aligned} Video\text{-}MR(f_i) = &\lambda Sim_1(f_i, V \setminus S) \\ &- (1-\lambda)\max_{g\in S} Sim_2(f_i, g) \end{aligned} \tag{3}$$

where V is the set of all frames in all videos, S is the current set of selected frames, g is a frame in S and f is a candidate frame for selection. Based on this measure, a summary can be constructed by selecting the keyframe iteratively with Video Maximal Marginal Relevance (Video-MMR):

$$\begin{aligned} S_{k+1} = S_k \cup \arg\max_{f_i \in V\setminus S_k} &(\lambda Sim_1(f_i, V \setminus S_k) - \\ &(1-\lambda)\max_{g\in S_k} Sim_2(f_i, g)) \end{aligned} \tag{4}$$

Where Sim_2 is the similarity between frames f_i and g. The keyframes are extracted with their current time in the video.

We identify the visual objects in the keywords [25] and bring the keywords information to the visual information extraction.

C. Text Processing

In this section, we process text, including speech transcripts, in order to extract the keywords from all the available text sources: video title, video description, and speech text. We also use the Neural Network text classification for brand safety and sentiment analysis to only keep user-friendly text content.

1) Text summarization: Text summarization [20] is the task of creating a short, accurate, and fluent summary of a longer text document. Basically, it contains two main tasks: keywords extraction and sentence extraction. In this paper, we apply the keywords extraction algorithm in order to extract the most representative keywords from all the text sources, for example, video title, video description, video speech text from Automatic Speech Recognition. The task of keyword extraction algorithm is to automatically identify a set of text keywords that best describe the whole text. The simplest possible approach is to use a frequency criterion with PageRank algorithm [21] after the word is converted to the vector. To implement text keyword summarization, we exploit the mature toolkit Gensim [22].

However, the keywords from text summarization contain the outliers in semantics, which are not representing the core meaning of the video text source. Thus, we propose to use a cross-validation filtering in text semantic meaning in order to purify the keywords to make them more consistent in the text meaning. We decide to use the words' cosine similarity to the other words in the keywords to filter out the outliers. We first convert the words to sparse vectors in high dimensions by word representation with Wikipedia trained data [24]. Then we compute the average cosine similarity of one word to all the other keywords according to Eq. 5.

$$sim_a = \sum_{n-1} sim_{a\in W, b\in W}(V_a, V_b))/(n-1) \tag{5}$$

where a and b are two words in keywords set W with the number of words n, and V is the vector of one word. If the sim_a is smaller than the selected threshold 0.15 from empirical experiments, we remove this word a from keywords cluster.

2) Text Classification: Text classification [10] is an important research topic of natural language processing (NLP), which is widely used in both research and industry with the popularity of artificial intelligence. To this end, we use a CNN model proposed by Yoon Kim[11] to do text classification for the purpose of defining brand safety and sentiment analysis.

a) Brand safety analysis: Brand safety is extensively noticed in the Internet business to avoid the unfavorable content or information being delivered to the customer. Brand safety cares the following domains according to normal definitions: adult, alcohol, arms, crime, death, hate speech, illegal drug, military conflict, obscenity, online piracy, spam, terrorism, and tobacco. We classify the text of a video by these 13 classes, and if the text is classified into one or multiple classes, the video skim would not be produced since the unfavorable content is detected.

b) Sentiment analysis: Simply speaking, Sentiment analysis [23] is to classify the text into two classes: positive (reader favorable) and negative (reader unfavorable). And we use the same approach of text classification [11] while consider sentiment analysis is a simplified problem of 2-classes classification with ”positive” and ”negative” compared to the normal text classification by CNN.

If one video is classified into one or multiple classes of brand safety or classified as ”negative”, we will not produce video skim of this video since it is not user-friendly.

D. MSI Information Fusion

In general, the application of data fusion methods contributes to those tasks that request any type of parameter estimation from multi-sources, which is an efficient method to regroup data. Information fusion refers to the process of integrating multiple sources to produce more consistent, accurate and useful information than that provided by any individual data source. We first review several relevant fusion methods in audio and video analysis and demonstrate our own fusion approach, which makes full use of the previous information.

1) Information Fusion Review: In an earlier literature[12], information fusion strategies have been classified into 3 levels: feature level, classifier level, and decision level. Feature fusion belongs to early fusion, which is the simplest to implement and is suitable for those applications which require very fast processing of data. However, it could not be applied to most tasks where strictly temporally synchronized data are not usable. Classifier fusion is one of the intermediate fusion strategies, which is an attempt to overcome the limitations of both early and late fusion strategies. In the classier fusion, we could apply a weighted combination of different modalities based on their reliability. These weighted combinations, however, are taken on each frame, allowing for a much ner combination of data than in late fusion. Such fusion schemes are widely used in audio-visual speech recognition systems.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 December 2018 doi:10.20944/preprints201812.0086.v1

The combination of probability scores or likelihood values obtained from separate uni-modal classifiers to come up with a combined decision are all involved in late or decision level fusion. The combinations of schemes with an appropriate weighting scheme has been used for audio-visual speech recognition. However, in the case of audio-visual speech recognition, the late fusion strategy has been shown to be inferior to the intermediate fusion strategy.

The proposed algorithm MSI tries to fuse all the potential indices by considering keywords relevance among them in order to ensure its semantic relevance.

2) MSI video skimming: In the above sections, we have explained in detail how we extract the following kinds of information from the video: 1) the keyframes; 2) the audio segments together with their audio classes; 3) the keywords with time stamps; 4) the user favorable decision by brand-safe and sentiment analysis. If the user favorable decision is positive, then we will first use time information to fuse all above kinds of information represented as keywords.

a) Time-Domain Information Fusion: The time domain information contains pitch time list and keyframes occurred time list. We regroup all these time point information as a long list, which contains the important time stamps in the video. Furthermore, we match speech text information to these time stamps and make a words list. Then we apply text summarization to this list of words. Besides, we apply also cross-validation to filter the keywords list. For all the summarized keywords, we re-match the time stamps and take the video keyframes or shots of these timestamps as one part the video skimming. As well we keep the keywords of these timestamps as a source of text summarized keywords to use in the next step. We illustrate the principle of time-domain information fusion in Fig. 2.

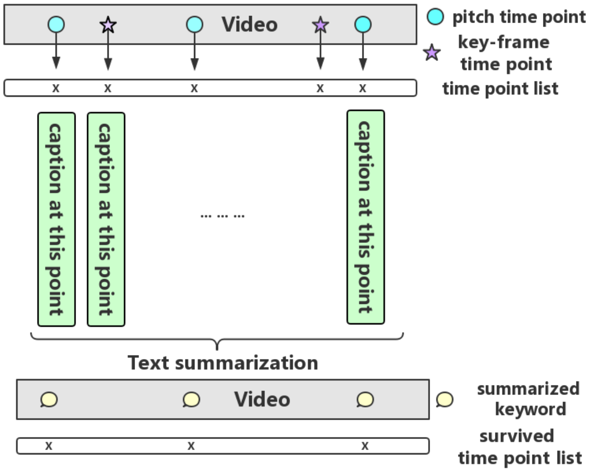

Fig. 2. Time-Domain Information Fusion

b) Keyword-Domain Information Fusion: With the text summarization step of all the sources ignoring the time information, we have extracted the keywords from the whole speech text with time stamps of text words. It is illustrated in Fig. 3.

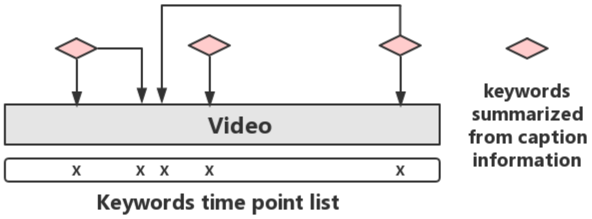

Fig. 3. Text Domain Information Fusion

By combining two above keywords with times stamps, we have obtained a complete time point list to generate the video skimming. Then we could apply keywords summarization and cross-validation filtering to remove the duplicated information again in order to get the final list of keywords with timestamps, which is illustrated in Fig. 4.

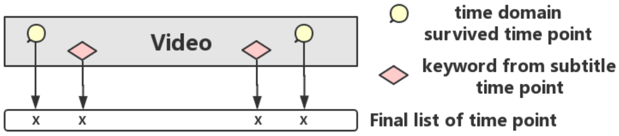
Video
time domain survived time point
keyword from subtitle time point
Final list of time point

Fig. 4. Combination of Two Time Point List

For each time stamp, we take this stamp as the center and form a fragment with the duration of each customized segment. Finally, the connected video segments are regarded as the final video skim produced by proposed MSI.

III. Experimental Results

We will describe the experimental data results in this section. Each information indices are in one subsection, with the followed data of text classification. MSI information fusion is in Section III-E.

A. Audio Segmentation

First, we generate a synthetic audio signal as an example to verify both the segmentation and classification model. The official label declaration is available in Fig. 9 and the audio signal is available[2]. In the result of the segmentation and classification test on the manually synthetic audio signal shown in Fig. 5, we check from the human level that the audio signal has been well-segmented. Then, for each segment, we apply the classification model to predict the class, if the audio segment is homogeneous, we save the prediction class result, for example, 0 denotes class speech, and 137 denotes class music. If not, for each part of the segment, we cut it into 10 seconds or shorter segments and also apply the classification model.

[2] https://drive.google.com/open?id=1k-eyIcTMTwRiVeq1D3X24UJS3lYozi8j

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 December 2018 doi:10.20944/preprints201812.0086.v1

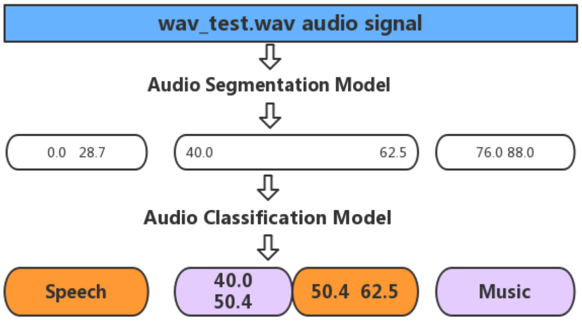

Fig. 5. Result of Segmentation and Classification on the Synthetic Audio Signal.

We had also tested with a special audio signal which contains pure and continuous gunshot sound as shown in Fig. 6. In this result, number 430 in Google Audioset denotes Artillery fire. After this test, we could find that the classification model could still recognize those weird sound like continuous gunshot sound.

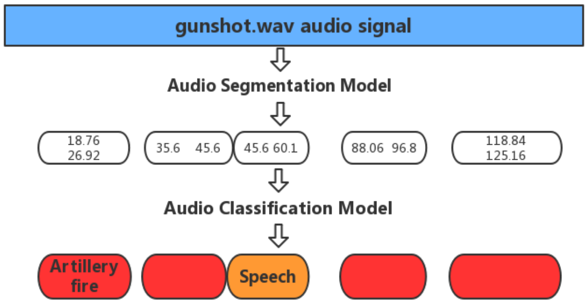

Fig. 6. Result of Segmentation and Classification on the Gunshot Audio Signal

Using this method of pitch detection, we test with several audio signals from YouTube platform, like the baby screaming in the zoo, video game recording with gunshot sound, movie clips, the pitch list detected is exactly the high-pitch point in the video. Applying first this segmentation model, we split the audio into small segments and then execute the classification step, which could save computation time significantly and some of the results will present in the following section.

B. Audio Classification

We choose Google AudioSet [26] as the training data of audio classification with MIL model, which is a large-scale collection of human-labeled 10-second sound clips drawn from YouTube videos. For each sound clip, 128-dimensional audio features extracted at 1Hz. The audio features were extracted using a VGG-inspired acoustic model, trained on a preliminary version of YouTube-8M [27]. The features are PCA-ed and quantized to be compatible with the audio features provided with YouTube-8M. They are stored as TensorFlow record files. Since the features provided by Google AudioSet is 128-dimensional audio embedding, we need to extract embedding features from the raw audio signal using VGGish model, which is a variant of the VGG model. It could be used as a feature extractor. VGGish model converts audio input features into a semantically meaningful, high-level 128 dimensions embedding which can be fed as input to a downstream classification model. The downstream model can be shallower than usual because the VGGish embedding is more semantically compact than raw audio features.

As the model is trained with Google AudioSet, which contains only 10s audio clips, to classify the long audio signal, we need to shift along the time sequence, the principle is showed in Fig. 7.

Each second, the raw audio was calculated 10 times, we used the predicted class most frequently seen in these 10 times as the final class for this second. In order to accelerate the speed of computing, we use GPU parallel computing to do the 10 times' prediction of 1-second audio duration. Having the embedding for the long audio signal, we convert them into a matrix to predict the label using the audio classication model pre-trained in order to shorten the executing time as shown in Fig. 8.

To verify the performance of the MIL classification model, we compare several models trained with different parameters, number of iterations different training corpus. Finally, test with a synthetic audio clip in 90 seconds, which contains two classes of audio signal, 'music' and 'speech'. The original class result i presents in Fig. 9. In Fig. 9 y=0 means class 'speech', y=137 means class 'music', and y=527 means class 'no label'. The X-axis presents the time, and the Y-axis presents the 528 classes provided by Google, from 0 to 526 are the original 527 classes. As we have mentioned previously, if the audio signal is recognized as no label, we defined its label as 527 for convenience, which means the output prediction of the model shows this audio clip belongs to none of the 527 originals classes. This synthetic audio signal in Fig. 9 contains speech and music part, 0-30s: speech, 30-50s: music, 50-60s: speech, 60-90s: music.

Since at the end of the training step, there is no more valuable improvement in accuracy, there might be the overfitting of the model. Obviously, using 50k iteration model there contains too much "no label" result and we could observe from the change point along the X-axis, the 30k iterations model has higher precision according to the raw audio signal. Therefore, in the following project, we use this 30k iterations model. The comparison of 30k and 50k iterations are shown in Fig. 9. The partial music section in 30K iteration model is recognized as "No label" because this part of music is not typical and could be forgiven after the human validation.

We also test the model with an audio signal, named as "scottish.wav" with original classes provided[13], which is a segment of radio recording contains human speaking and music. The ofcial label and predicted label using pre-trained 30k iterations model are in the figure above. Although the result contains several "no label" prediction, the overall trend

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 December 2018 doi:10.20944/preprints201812.0086.v1

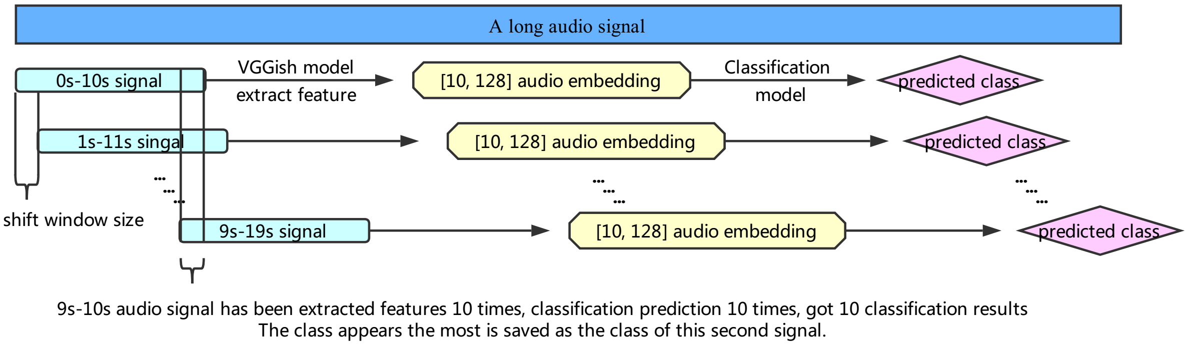
A long audio signal
0s-10s signal
VGGish model
extract feature
[10, 128] audio embedding
Classification
model
predicted class
1s-11s singal
[10, 128] audio embedding
predicted class
shift window size
9s-19s signal
[10, 128] audio embedding
predicted class
9s-10s audio signal has been extracted features 10 times, classification prediction 10 times, got 10 classification results
The class appears the most is saved as the class of this second signal.

Fig. 7. Scheme for Audio Classification

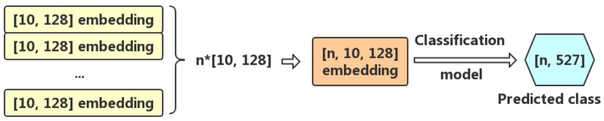

Fig. 8. Scheme for Matrix Computation

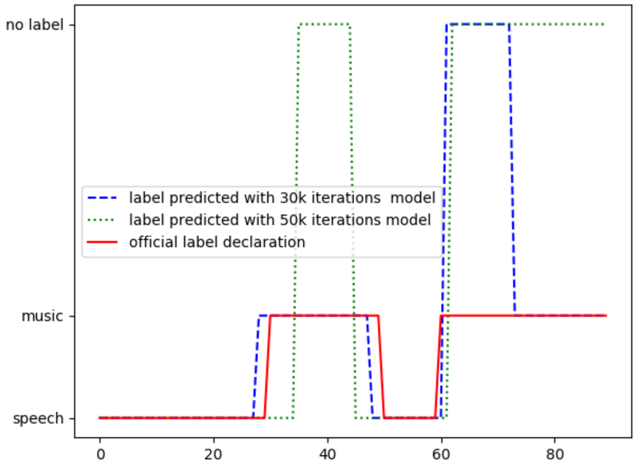

Fig. 9. Synthetic Audio File Official Label Declaration and Predicted Label with Different Models

of classification is correct and speech in music is reasonable to be recognized as ”speech”, as shown in Fig. 10.

We could conclude that the model is useful. However, the class prediction has about 2-5 seconds deviation with the original label, which means it could not exactly get the audio class change point, which might be caused by its 10 seconds features. In fact, it is not influencing our approach whether we could recognize the change point of the audio signal class in our following discussion since we only consider the class exists along the audio signal and pay less attention to the change point of the classification result of every single second. Under this condition, we apply the audio classification model in the following work.

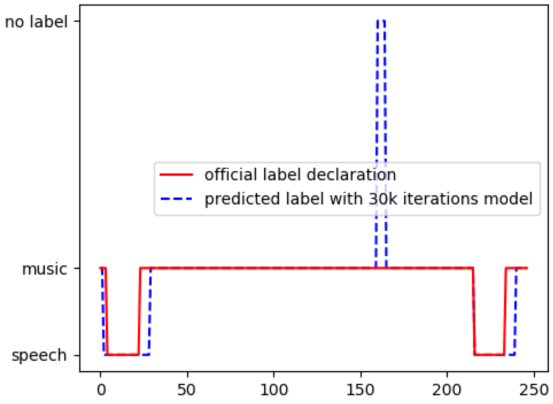

Fig. 10. 30k Iterations Model Evaluating with scottish.wav Audio Signal

The recognized audio class name is used as keywords too in video skimming based on the keywords.

C. Keyframe extraction

In our experiment, the extracted keyframes successfully represent the important video scenes and contents. In Fig. 11 We show it as an example keyframe of one selected video[3].

D. Text classification

To train the sentiment analysis model, we use Internet Movie Database corpus[4] with CNN sentence classification [28]. While for brand safety, we set 13 classes to as”adult”,

[3] Video chosen from Youtube platform: https://www.youtube.com/watch?v=nqnkBdExjws

[4] https://www.imdb.com/interfaces/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 December 2018 doi:10.20944/preprints201812.0086.v1

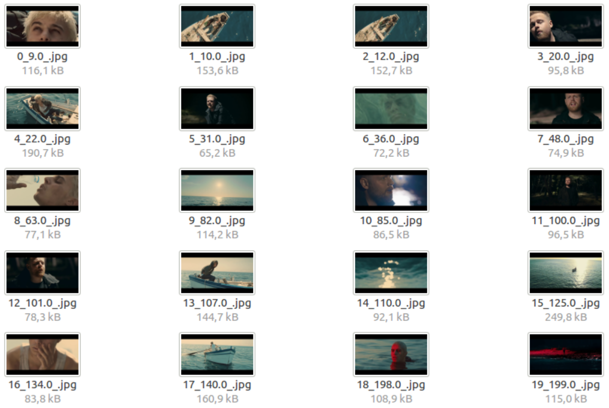

Fig. 11. Keyframes Extract Result Using Video-MMR, Youtube ID='nqnkBdExjws'.

"alcohol", "arms", "crime", "death", "hate speech", "illegal drug", "military conflict", "obscenity", "online piracy", "spam", "terrorism", and "tobacco". The training data of brand safety is collected from Internet.

```
text = [
'The baby is lovely. I like him.',
'The drug is dangerous for the world.',
'Everyone should have the gun in the war.',
'I like women.']
classify_result = [[],['crime'],['arms'],[]]
```

Fig. 12. Example for Text Classification

From this example, we could easily discover that the text classification is able to detect adult issue from sentences. According to our test, we could basically conclude that the text classication model performs well and have achieved the similar accuracy described in [28], 88.7%. In addition, the training and test corpus are collected by web crawling.

E. MSI Information Fusion

To valid this video skimming generation method, we chose the video from the YouTube-8M dataset [27] with the news class. To generate the video skim, we use the speech text from the video which has been provided in YouTube in English. Firstly, to choose the video from YouTube-8M dataset, the reason why we choose the video in news class is these videos always make up of several different scenes, which is convenient for us to check the video skimming quality. The other condition is to select the video with English caption from speech recognition. Under these two requirements, we selected several videos for the test.

After choosing the list of video IDs for the validation of the method[5], we generate the video skim for each of them and verify the proposed approach from multiple perspectives:

1) User-pleasant quality: Removing those video segments which contain adult issue or dangerous words like 'gun','sex' and so on, with the result from text classification in both audio and keyframes time-point caption information, we finally check the video skim whether they still include these information. According to the 19 video skims we've generated, none of them contains these words, which means they are the truly user-pleasant result.

2) Semantic quality: After checking the visual quality of the video skims, we turn to verify the semantic quality by keywords. Since the text information we have extracted from the previous analysis is the keywords lists, we tried to calculate the similarity score between keywords representing the video skim and all the text information from the video to check whether the video skim is meaningful and representative from the semantic level. To the best of our knowledge, it is a novel approach and evaluation for video skim in keyword semantics.

We prepare the following 3 kinds of text keywords from video skims: 1) For each video, we get the complete speech text and summarize keywords from it, which we called text_keyword. This is purely from text information of the video 2) From the keyframes we extract using the Video-MMR algorithm, we also have a list of keyframe timestamps, and then extract the speech text with these time stamps and get the keywords, which we called text_keywords. This is the keywords from visual information of the video. 3) The video skim is generating by the proposed MIL approach, which we called MSI_keyword from multi-model information. Having these three lists of keywords, we use the similarity model to calculate the similarity score between subtitle_keywords list and the other two lists. Here is the equation that presents how we calculate the similarity score between two lists of words.

$$Sim\_score\_video\_skim = \frac{1}{N}(\max S_{W_i,L_j}) \qquad (6)$$

where $j = 1...N$, W_i is the word in the first keywords list, and L_i is the word in the second list, N is the length of this second list.

In this way, we obtain three similarity scores of three keywords list with the same keywords number comparing to the text information of the video, which could explain how better the information in the keywords list presents the main content of the original video caption. The results are shown in Fig. 13. We use these similarity scores as the evaluation. Finally, we found the Sim_{MSI} is always much higher than the Sim_{visual} and the Sim_{text}, where Sim_{MSI} is 37.5% higher than Sim_{visual} in average and 44.3% higher than Sim_{text} in average. Thus we conclude that the video skims from the proposed MSI approach have well summarized the semantic content of the video in the universe of keywords semantics and outperforms the approaches only using text keywords or visual information.

IV. CONCLUSION

In this paper, we have proposed a MSI video skimming generation approach, which makes full use of the original

[5] Original videos chosen from Youtube-8M dataset: https://docs.google.com/document/d/1Wi55nPHvTIjgkMY-vGXDLexhvdL3cRn99aZ1mCNmwkM/edit?usp=sharing

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 6 December 2018 doi:10.20944/preprints201812.0086.v1

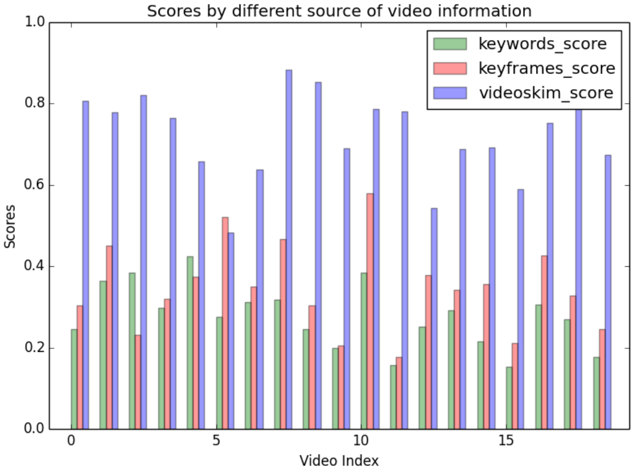

Fig. 13. Similarity Score Comparison

information of the video and try to conserve semantic content of the video by temporal keywords relevance. In order to extract useful information in each video information channel, we exploit the state-of-the-art deep learning approaches. Then we regroup and fuse semantic contents represented by temporal keywords relevance. To keep away from the unfavorable contents like the adult content, we exploit the brand-safety and sentiment classification to detect unfavorable issues supported by the audio and text classification. Finally in the experiment, we compare the video skims using different approaches with respect to keywords semantics while video skim generated by MSI has better performance than video skims generated only by text or visual information. In the future, we intend to improve the fluency of video skimming, for example, avoiding semantic breaking in the skimming.

References

[1] Yu-Fei Ma and Hong-Jiang Zhang, A model of motion attention for video skimming, Proceedings. International Conference on Image Processing, Rochester, NY, USA, 2002, pp. I-I.

[2] Qiuqiang Kong, Yong Xu, Wenwu Wang, and Mark D. Plumbley. Audio set classification with attention model: A probabilistic perspective. CoRR, abs/1711.00927, 2017.

[3] M. A. Smith and T. Kanade, Video skimming and characterization through the combination of image and language understanding, Proceedings 1998 IEEE International Workshop on Content-Based Access of Image and Video Database, Bombay, India, 1998, pp. 61-70.

[4] N. Dimitrova, H. Zhang, B. Shahraray, I. Sezan, T. Huang and A. Zakhor, ”Applications of Video-Content Analysis and Retrieval,” in IEEE MultiMedia, vol. 9, no. 3 pp. 42-55, 2002.

[5] T. Liu et al., ”Learning to Detect a Salient Object,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 33, no. 2, pp. 353-367, Feb. 2011.

[6] Hannane, R., Elboushaki, A., Afdel, K. et al. Int J Multimed Info Retr (2016) 5: 89. https://doi.org/10.1007/s13735-016-0095-6

[7] Daniel Fallman. The penguin: Using the web as a database for descriptive and dynamic grammar and spell checking. In CHI 02 Extended Abstracts on Human Factors in Computing Systems, CHI EA 02, pages 616617, New York, NY, USA, 2002. ACM

[8] Jorey Ramer, Adam Soroca, and Dennis Doughty. Predictive text completion for a mobile communication facility, March 15 2007. US Patent App. 11/422,797.

[9] James Allen. Topic Detection and Tracking: Event-based Information Organization. Kluwer Academic, 2002.

[10] Chen Wenliang, Zhu Jingbo, Zhu Muhua, and Yao Tianshun. Text representation using domain dictionary [j]. Journal of Computer Research and Development, 12:019, 2005.

[11] Yoon Kim. Convolutional neural networks for sentence classification. CoRR, abs/1408.5882, 2014.

[12] Claire B, Riccardo B, Marco V, et al. Audiovisual Information Fusion in HumanComputer Interfaces and Intelligent Environments: A Survey[J]. Proceedings of the IEEE, 98(10):1692-1715, 2010.

[13] Theodoros Giannakopoulos. pyaudioanalysis: An open-source python library for audio signal analysis. PLOS ONE, 10(12):117, 12 2015

[14] Li, Yingbo, and Bernard Merialdo. ”Multi-video summarization based on Video-MMR.” Image Analysis for Multimedia Interactive Services (WIAMIS), 2010 11th International Workshop on. IEEE, 2010.

[15] Lu, Lie, Hong-Jiang Zhang, and Stan Z. Li. ”Content-based audio classification and segmentation by using support vector machines.” Multimedia systems 8.6 (2003): 482-492.

[16] Lee, Honglak, et al. ”Unsupervised feature learning for audio classification using convolutional deep belief networks.” Advances in neural information processing systems. 2009.

[17] Lew, Michael S., et al. ”Content-based multimedia information retrieval: State of the art and challenges.” ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 2.1 (2006): 1-19.

[18] Ejaz, Naveed, Tayyab Bin Tariq, and Sung Wook Baik. ”Adaptive key frame extraction for video summarization using an aggregation mechanism.” Journal of Visual Communication and Image Representation 23.7 (2012): 1031-1040.

[19] Mundur, P., Rao, Y. , Yesha, Y. Int J Digit Libr (2006) 6: 219. https://doi.org/10.1007/s00799-005-0129-9

[20] Allahyari, Mehdi, et al. ”Text summarization techniques: a brief survey.” arXiv preprint arXiv:1707.02268 (2017).

[21] Page, Lawrence, et al. The PageRank citation ranking: Bringing order to the web. Stanford InfoLab, 1999.

[22] Rehurek, Radim, and Petr Sojka. ”Software framework for topic modelling with large corpora.” In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. 2010.

[23] Pang, Bo, and Lillian Lee. ”Opinion mining and sentiment analysis.” Foundations and Trends in Information Retrieval 2.12 (2008): 1-135.

[24] Pennington, J., Socher, R., Manning, C. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

[25] Redmon, Joseph, and Ali Farhadi. ”Yolov3: An incremental improvement.” arXiv preprint arXiv:1804.02767 (2018).

[26] https://research.google.com/audioset

[27] Abu-El-Haija, Sami, et al. ”Youtube-8m: A large-scale video classification benchmark.” arXiv preprint arXiv:1609.08675 (2016).

[28] Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882.

[29] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. ”Imagenet classification with deep convolutional neural networks.” Advances in neural information processing systems. 2012.