A Hybrid Approach to News Video Classification with Multi-modal Features

Peng Wang†, Rui Cai and Shi-Qiang Yang

Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
†Email: wangp01@mails.tsinghua.edu.cn

Abstract

This paper presents a hybrid approach to the classification of news video story. Most of current works on news story classification utilize the multi-modal features in a uniform manner. However, the reliability of audio-visual confidence is much lower than that of text, which may evidently lower-down the performance of the classification. We proposed a decision strategy mainly depends on the evidence from text classifiers with extra assistance of audio-visual clues. In our approach, SVMs for text features and GMMs for audio-visual features are first built for each category and then used to compute text and audio-visual confidence vectors respectively. To make final decision, a text-biased decision strategy is proposed to combine these multi-modal confidence vectors. To validate the performance, text-based classification and SVM-based meta-classification methods are compared on large-scale news stories from TV programs, and our proposed hybrid approach achieves the best overall performance.

1. Introduction

To classify single-story news video into conceptual categories will dramatically facilitate users' browsing in a flexible way and provide the possibility for users to retrieve the target news video. Semantic classification is the fundamental problem of content-based retrieval researches and applications. Tremendous efforts have been made to segment broadcasting program into single-story units and assign each story to a specific domain. In [1], a statistical approach based on closed-captioned text was proposed for news video segmentation and categorization. Liu et al. [2] proposed a technique to identify different audio events in broadcasting video using Hidden Markov Models with the help of audio cues. Since a single modality does not provide sufficient information for accurate multimedia content classification, combining evidence from multiple modalities for video classification has been investigated to improve categorization accuracy in several researches. In [3], they integrated image and audio analysis in the identification of news segments, and applied video OCR technique to acquire textual information for story classification. The work in [4] introduced a meta-classification combination strategy with SVM integrating evidence from multiple modalities. Multi-modal features are useful for the news video semantic classification. How to fuse them in classification is still an open problem. However, most of existing works on news story classification utilize the multi-modal features in a uniform manner. It is difficult to obtain reliable news story classification based on audio-visual cues because of the baffle of current techniques of image understanding and audio signal processing. Nevertheless, text categorization has been well-studied for several decades and achieves high performance in practice [5]. Thus text information is more reliable and effective than audio-visual features in news video classification. In this paper, a hybrid approach is proposed to classify news video in a text-biased manner.

In the hybrid approach, the decision strategies mainly depend on text information with extra assistance of audio-visual clues. In this approach, Support Vector Machines are exploited to compute text confidence scores from low-level textual features and Gaussian Mixture Models are employed to compute audio-visual confidence scores from low-level audio-visual features. The decision strategies treat the multi-modal confidence scores as input and perform the classification in a text-biased way. We apply the SVM-based meta-classification combination strategy [4] as a consultant in our proposed decision strategies. Ten story categories are defined in current experiments: Politics(P), Military(M), Sports(S), Weather(W), Transport(T), Business(B), Health(H), Entertainment(E), Science&Technology(C) and Daily(D), which covers most species of news video. The proposed approach has been validated by the experiments on large-scale living video programs.

The rest of this paper proceeds as follows. In section 2, we describe the extraction of textual, audio and visual features. Section 3 presents the hybrid approach to automatic news video categorization in detail. The experiments and evaluation are shown in section 4. Finally, the conclusion and discussion are given in section 5.

2. Multi-modal Feature Extraction

In our system, three types of features are extracted at different granularities: (i) textual features from one story; (ii) audio features from one-second clip; (iii) visual features from one shot. These selected granularities are the basic units which widely used in corresponding domains for low level feature analysis.
2.1 Textual Feature Extraction

Because the transcripts are usually unavailable in news video analysis, we use a speech recognizer [11] to produce transcripts of the spoken text (ToS). Since most speech in news video is very clear and exact, we could achieve high recognition accuracy above 90%, which provides effective cues for advanced classification. For each story, the term-frequency and inverse-document-frequency (TF-IDF) features are extracted [5].

\( TF(t, w) \) represents the number of times that term \( w \) occurs in text \( t \), and is known to be used to improve the recall of text classification. \( IDF(w) \) describes the specificity of term \( w \) and is able to improve the precision. It is defined as:

\[
IDF(w) = \log(N / n(t, w))
\]

where \( N \) is the total number of texts in the database and \( n(t, w) \) is the number of text \( t \) which contains the term \( w \).

To avoid unnecessarily large dimension of the feature vector, terms are considered only if they occur in the training data with adequate times no less than a pre-defined threshold (3 times in our experimental system). Each story is denoted by the TF-IDF vector:

\[
ft = (t_1, t_2, \cdots, t_n)
\]

where \( n_1 \) is the dimension of \( ft \), which equals 304 in our implementation.

2.2 Audio-Visual Feature Extraction

As an important modality, audio analysis has been performed in fields of audio events identification [2] and audio classification [6]. Grounded on the previous works, the audio track of news video is first down-sampled to 8 kHz, and then each one-second clip is divided into 40 non-overlapping 25ms-frames. For each audio frame, a set of temporal and spectral features, which have been proved to be useful in discrimination of speech/music/environment [6] are extracted to help identifying news categories. These features include high zero-crossing rate ratio, low short-time energy ratio, brightness, bandwidth, spectrum centroid, spectrum rolloff, sub-band energy, 8-order MFCC and pitch. Statistical characters of these frame-based features, which include mean, standard deviation and autocorrelation of a small lag, are calculated over clip and are connected with some clip-based features, such as spectrum flux, band periodicity and noise frame ratio to form the audio feature vector, each clip is denoted as:

\[
fa = (a_1, a_2, \cdots, a_n)
\]

where \( n_2 \) is dimension of \( fa \), which equals 49 in our implementation.

The visual stream is automatically segmented into shots and corresponding key frames are also extracted. The selected visual features are essential to model the semantic contents of shots, including face, closed-caption, black frame, shot duration and motion energy. For face, we adopted the latest Intel OpenCV library [12] to detect the number of mostly frontal faces and their sizes in the key frame. For closed-caption, we apply the method developed in [7] to detect the number of lines and size of videotext appearing in the key frame. With the method in [8], we extracted the motion energy features in frame sequence, and their mean and variation values are calculated for each shot. By combining the above features into one vector, each shot is denoted as:

\[
f_v = (v_1, v_2, \cdots, v_n)
\]

where \( n_3 \) is the dimension of \( f_v \), which equals 14 in our implementation.

3. Hybrid Decision Strategies

3.1 Confidence Computation

3.1.1. Text Confidence Vector

SVMs are exploited to compute confidence from low-level text features in our approach, which have been validated in text categorization [5]. We construct \( \mathcal{K} \) 2-class SVM models for \( \mathcal{K} \) news story categories respectively, where Gaussian Radial Basis function is chosen as the kernel of SVMs. The \( k \)th SVM model is trained with all stories in the \( k \)th category with positive labels, and all other stories with negative labels. To obtain confidence score of each story to each category, calibrated probabilistic outputs of SVMs are required; however, standard SVMs do not provide such output. Platt [9] proposed a method by training the parameters of an additional sigmoid function to resolve this issue. The LIBSVM software [10] used in our experiments implements this method and provides the very function to enable us to generate confidence vector in text categorization. For a given news story, one \( \mathcal{K} \)-dimensional confidence vector will be obtained after its text feature \( ft \) is tested through all the \( \mathcal{K} \) SVMs. In order to facilitate post processing, we convert the probabilistic output of SVMs to logarithmic coordinate, and the text confidence vector can be represented as:

\[
CT = (ct_1, k = 1, \cdots, \mathcal{K})
\]

where \( ct_k \) is the text confidence to the \( k \)th category.

3.1.2. Audio-Visual Confidence Vector

Gaussian Mixture Models (GMMs) are used to model the low-level audio feature and map them to confidence vector. In our approach, \( \mathcal{K} \) GMM models are trained to describe the audio features of \( \mathcal{K} \) story categories respectively. The \( k \)th GMM is trained with all the stories in the \( k \)th category. The component number of Gaussian mixture is selected as 64 and a standard EM algorithm is applied to estimate the
parameters. Let the trained $k^{th}$ GMM be denoted as $G_k(\omega_k, m_k, \Sigma_k)$ ($0 \leq i < 64$), with $\omega_k, m_k, \Sigma_k$ being the weight, mean and deviation of each Gaussian component.

For a testing story, suppose the extracted audio features are

$$FA = \{f_{a1}, \ldots, f_{aM}\}$$

where $N$ is the number of the segmented audio clips in this news story. Then log-likelihood score of $k^{th}$ clip to $k^{th}$ audio GMM can be calculated by:

$$p_i(f_{ai} | G_k) = \sum_{j=1}^{M} \omega_k p_i(f_{ai})$$

and

$$p_i(f_{ai}) = \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \text{exp}\{-\frac{(f_{ai} - m_i)^T \Sigma^{-1} (f_{ai} - m_i)}{2}\}$$

The confidence score of the story to $k^{th}$ GMM is calculated as:

$$ca_k = \frac{1}{N} \sum_{i=1}^{N} p_i(f_{ai} | G_k)$$

and the audio confidence vector of the news story is denoted as:

$$CA = (ca_k, k = 1, \ldots, K)$$

Similar with audio feature, GMMs is also applied to model the low-level visual features and compute visual confidence vector. The training and testing of visual GMMs are done in the same way as that of audio GMMs. For a given news story, the extracted visual features are

$$FV = \{f_{v1}, \ldots, f_{vM}\}$$

where $M$ is the number of detected shots in this story. We can obtain $p(f_{vi} | G_k)$, which is the likelihood score of the $i^{th}$ shot to the $k^{th}$ visual GMM ($G_k$). Then, the visual confidence score of the news story to $k^{th}$ GMM is calculated as:

$$cv_k = \frac{1}{N} \sum_{i=1}^{N} \sum_{i=1}^{M} w_i \cdot p_i(f_{vi} | G_k)$$

where $w_i$ is the weight of the $i^{th}$ shot, which is determined by its shot duration in our experiments. And the visual confidence vector can be represented as:

$$CV = (cv_k, k = 1, \ldots, K)$$

### 3.2 Combination Strategies

The combination strategies determine how the information from multi-modal confidence vectors is combined together to do the final categorization. In this section, we describe a meta-classification combination strategy using SVMs and propose our text-biased strategies that textual confidence vector has a prior decision than that of audio-visual confidence.

#### 3.2.1. SVM-based Meta-classification

Lin et al. [4] proposed an SVM-based meta-classification combination strategy, which treats the judgment from each classifier for each class as a feature, and build another classifier, i.e. a meta-classifier, to make the final decision. We take the calculated confidence vectors of textual, audio and visual features as the judgment from multiple classifiers, so that these vectors are combined into a long feature vector $\mathbf{f} = (ct_1, \ldots, ct_K, ca_1, \ldots, ca_K, cv_1, \ldots, cv_K)$ as the input of the meta-classifier. The meta-classifier takes the feature vector $\mathbf{f}$ and makes a final decision. As suggested in [4], SVM can be chosen as the meta-classifier. A $K$-class SVM is built here to achieve final decision.

#### 3.2.2. Text-biased Strategies

The reliability of audio-visual confidence is much lower than that of text in the case of news video classification, which may evidently lower-down the performance of the SVM meta-classification. Therefore, we propose the text-biased strategies to classify the video using the multi-modal confidence vector. As shown in Figure 1, the decision strategies are composed of two phases. Firstly, if $CT$ is discriminable, a decided category is produced; otherwise a set of candidate categories (CC) is constructed based on the $CT$. Secondly, a decided category was estimated from the candidate classes, where the SVM-based meta-classification is employed as a consultant.

![Flow diagram of the decision method for news video categorization](image)

In the first step, $CT = (ct_k, k = 1, \ldots, K)$ are compared one another and sorted in descending order, which is represented as $CT = \{ct_1, ct_2, \ldots, ct_K\}$. This condition is adopted to check whether the maximum confidence is higher enough than others, which ensures the reliability and accuracy of the decided category. If the condition is not satisfied, the set of candidate categories will be constructed as follows.

**Step1.** Calculate $c = \text{arg}_k (ct_k = ct'_1)$, then set candidate classes $CC = \{c\}$;
Step2. Calculate $c = \arg_k (c_t = c_{t_2})$, then set $CC = CC \cup \{c\}$. Set variable $j = 3$;

Step3. Calculate $\Delta = c_{t_j} - c_{t_{j-1}}$, if condition ($\Delta > T_2$) is satisfied, goto Step 5, where $T_2$ is the same with that of the previous condition;

Step4. Calculate $c = \arg_k (c_t = c_{t_j})$, then set $CC = CC \cup \{c\}$, set $j = j + 1$; If condition ($j < K$) is satisfied, goto Step 2;

Step5. Finish constructing candidate categories $CC$.

In the second step, to estimate a decided category from the candidate classes, the estimation of an SVM-based meta-classification is consulted. Here, we have implement the task from the $CT$, $CA$ and $CV$ of the input news story. If the estimation of meta-classification $C_k$ belongs to the set of candidate classes $CC$, we confirm it and make $C_k$ as the final decision; otherwise we will still take the class with the maximum text confidence in $CC$ as the final decision.

Compared with the SVM-based meta-classification strategy, our text-biased decision strategies mainly depend on text confidence vector because text has the prior capability in the task of news video categorization. Although the SVM-based meta-classifier has attempted to learn the weights for different classifiers, the unreliable outputs of audio and visual classifiers possibly weaken its performance. From this point of view, when text confidence is discriminative enough, we can make the decision without considering audio-visual confidence vectors, which possible confuse the decision; otherwise, the advantage of the SVM-based meta-classification will be exploited in making decision from the candidate classes. In this way, the most reliable text classifier is considered sufficiently as well as taking the advantage of the SVM-based meta-classification in combining multi-modal classifiers.

4. Experiments

To evaluate the effectiveness of the proposed method, experiments on 559 news stories from CCTV of China are carried out. These stories have been manually segmented and labeled for both training and testing purpose. 30 stories (about 50%) are chosen randomly for each category as the training data and the remainders are used as the testing data.

First, those news videos are classified only based on textual features, which classify the story to the category having the maximum text confidence score. As shown in Table 1, we obtain the recall rate of 62.4% and precision rate of 88.4% on average. Second, using the SVM-based meta-classification strategy as described in section 3.2.1, and repeating the experiments, we obtain the recall rate of 79.1% with precision rate of 78.9% for average classification.

The experimental results of our proposed text-biased method are given in Table 1 with two different methods as references. Compared with the former two methods, text-biased hybrid classification method proposed has the best overall performance with the recall rate of 77.8% and precision rate of 84.6%. This test shows that although in theory SVM-meta-classifiers have the best performance and our proposed method is superior in practice because of the baffling of current image and audio analysis techniques.

<table>
<thead>
<tr>
<th></th>
<th>Text-based</th>
<th>SVM-Based</th>
<th>Two-step</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>62.4%</td>
<td>79.1%</td>
<td>77.8%</td>
</tr>
<tr>
<td>Precision</td>
<td>88.4%</td>
<td>78.9%</td>
<td>84.6%</td>
</tr>
</tbody>
</table>

Figure 2 illustrates the performance comparison of these three methods.

- The categories of Sports and Weather have the most salient textual features and do not easily confuse with any other categories. They achieve both high recall and precision rates no matter which method is used.

- The categories of Politics, Military, etc, except for Daily have relatively salient textual features, but easily confuse with Daily. They achieve high precision but low recall rates using the text-based method. Although SVM-based meta-classification increases their recall rates dramatically, their precision rates are decreased greatly also. Our text-biased method obtains both moderate precision and recall rates.

- The category of Daily is a very broad domain without unique textual features, and many other categories are easily classified into it by mistake. It achieves high recall but very low precision rate using text-based method. SVM-based meta-classification increases its precision rate largely, but decreases the recall rate also. However, our proposed text-biased approach keeps both moderate precision and recall rates.
5. Conclusions

In this paper, a novel approach for news story classification is presented, which is hybrid combining text-biased classification and SVM-based meta-classification. The proposed combination strategies mainly depend on text confidence vector with extra assistance of audio-visual confidence vectors. SVMs and GMMs are respectively exploited to compute multi-modal confidence vectors. The hybrid combination strategies combine the multi-modal confidence vectors to making final decision. We have compared our approach with text-based classification and the SVM-based meta-classification in experiments. The results show that the proposed text-biased strategies perform better than the other two methods. In addition, other video analysis technologies can be applied to enhance the user-oriented video browsing, for example user browsing behavior can be analyzed to optimize the organization of video classification. This will be the focus of our future work.

References


