International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04), Montreal, Canada, May 17-21, 2004

Multi-modal Classification in Digital News Libraries

Ming-yu Chen
School of Computer Science
Carnegie Mellon University
Pittsburgh PA, USA 15213
+1 412 268 7003
mychen@cs.cmu.edu

Alexander Hauptmann
School of Computer Science
Carnegie Mellon University
Pittsburgh PA, USA 15213
+1 412 268 1448
alex@cs.cmu.edu

ABSTRACT
This paper describes a comprehensive approach to construct robust multi-modal video classification on a specific digital source, broadcast news. Broadcast news has a very stable structure and every segment has its specific purpose. Video classification can support fundamental understanding of the structure of the video and the content. The variety of video content makes it hard to classify; however, it also provides multimodal information. Our approach tries to solve two important issues of multimodal classification. The first one is to select few discriminative features from many raw features and the second one is to efficiently combine multiple sources. We applied Fisher's Linear Discriminant (FLD) for feature selection and concatenated the projections into a single synthesized feature vector as the combination strategy. Experimental results on the 2003 TRECVID [1] news video archive show that our approach achieves very robust and accurate performance.

Categories and Subject Descriptors
H.5.1 [**Multimedia Information Systems**]: Video.

General Terms
Algorithms

1. INTRODUCTION
Broadcast news archives are a valuable and abundant source of content for digital libraries. To manage this large information archive, retrieval and categorization are two essential technologies. However, they all rely on understanding the content and structure of news videos. Although it is abundant and varied, news video has stable structure and meaningful segments. Normally, a news video will contain anchorperson, news event, commercial, weather report, and sports report segments. To detect and classify video as containing these scenes provides intermediate knowledge helpful to analyze and understand the content. The basic idea of video classification is to extract features from different sources (e.g. images, audio, transcript and speech) and to classify based on those features. We propose a comprehensive approach to solve two major problems in video classification. First, we applied Fisher's Linear Discriminant (FLD) to select features. Second, we synthesized a single new vector from the result of FLD to represent the content of the scene. Based on this representation, we applied Support Vector Machine (SVM) as classifier. We tested our approach by constructing anchor and commercial detectors for the TREC 2003 Video Track (TRECVID). The experimental results show that our approach is robust and can achieve high accurate performance.

JCDL'04, June 7–11, 2004, Tucson, Arizona, USA.

2. FEATURE SELECTION
Fisher's Linear Discriminant is a useful feature selection method. It is a classification method that projects high-dimensional data onto a lower dimensional space. The projection maximizes the distance between the means of classes while minimizing the variance within each class. This defines the Fisher criterion:

$$S_w = \sum_{i=1}^{c} \sum_{x_k \in C_i} (x_k - u_i)(x_k - u_i)^T \quad (1)$$

$$S_b = \sum_{i=1}^{c} |C_i|(u_i - u)(u_i - u)^T \quad (2)$$

$$J(w) = \frac{w^T S_b w}{w^T S_w w} \quad (3)$$

where S_b is the distance between each class, S_w is the variance within each class, and w is the possible projection. The optimal projection w_{opt} is:

$$w_{opt} = \arg\max_w J(w) \quad (4)$$

From the Lagrange Multiplier Rule, we find that

$$S_b w = \lambda S_w w \quad (5)$$

This is equivalent to solving a generalized eigenvalue problem. The optimal w is the eigenvector corresponding to the largest eigenvalue of the equation 5.

The projection can be viewed as giving a weight to each dimension and making the data more discriminable. Figure 1 shows the weights resulting from applying FLD on a color histogram for anchor detection. High weights on grids 2, 5, 17 and 20 indicate the studio background as an important clue to detect anchor scenes.

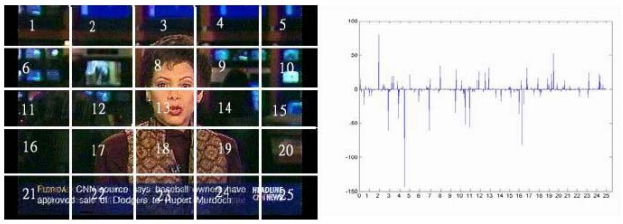

Figure 1. FLD weights for anchor detection. The image grid is on the left, the right graph shows the weight of every grid.

Category	MAP Anchor	MAP Commercial
Image Features only	0.71	0.83
Face Information only	0.83	N.A.
Speaker ID/Audio only	0.64	0.67
Feature Synthesis	0.79	0.85
Meta-classifier	0.83	0.88
FLD Synthesis	0.87	0.93

Table 1. Anchor and Commercial classifier results. MAP is mean average precision

3. COMBINATION STRATEGY

There are generally two strategies for combining features: feature synthesis which merges different kinds of features into one integrated vector; or classifying different feature sets and then combining the classification results into a final decision. Feature synthesis tries to represent the content of multiple media as one integrated feature vector. It is a simple idea and an intuitive way to do the combination, but most experiments show that it does not work well. The second approach classifies every feature set individually and then combines the classified results. This approach tries to simplify the content of multiple media and assign a higher level meaning to each media feature set by applying a binary classifier judgment to every set. We can then make discriminative decisions based on these judgments. The main drawback of this strategy is that detail information contained in the feature sets is lost in the process of shrinking the dimensionality of each set to one classifier result or judgment.

The basic idea of our combination approach is to obtain the benefits of both combination ideas. We apply FLD to every feature set and synthesize new feature vectors from every set. This step can be interpreted in two ways. First, it is feature selection. Second, it is classification of the data since FLD's target function has the inherent ability to discriminate between classes of data. New feature vectors are not only selected from the raw data, but also generated by a discriminant function. Based on those new feature vectors, we synthesize a single concatenated feature vector to represent all the multimedia content and then apply classification using this representation.

4. EXPERIMENTAL RESULTS

Our experiments are carried out on a subset of data provided by TRECVID 2003. 12 CNN news shows were randomly selected as our corpus and correct shot classification was manually determined. The common shot segmentations, defined by TRECVID, were used as the basic units. One keyframe was extracted for each shot.

We developed two detectors based on our approach. The anchor detector which takes advantage of image features, face information and speaker ID as feature sets. **Image features** are based on a 5-by-5 125-bin color histogram. **Face information** contains size, position, and confidence predicted by a face detector [2]. **Speaker ID** was provided by LIMSI [5]. The commercial detector utilizes image features and audio features. **Audio features** are Short Time Fourier Transform. 5-folder cross validation was performed to evaluate performance. The SVM classifier we used was LIBSVM [3].

Feature synthesis is the method to merge feature sets as one integrated feature vector. **Meta-classifier** [4] is the method that builds a classifier on the classified result of each feature set. **FLD synthesis** is the approach we proposed. It applies FLD to project high-dimensional features to lower dimensions and concatenates the FLD result of each set into one integrated feature vector. Experimental results in Table 1 show that FLD synthesis outperforms the other two combination strategies. It also shows that multimodal performs better than single-modal classification.

5. CONCLUSIONS AND FUTURE WORK

Feature selection and combination strategy of multimodal information are two important issues in constructing a video classifier. We propose a comprehensive approach to build a robust multiple media classifier. FLD is applied to select features and FLD synthesis performs well as a combination method. Results from anchor and commercial detection experiments on the TRECVID 2003 archive prove that our approach achieves robust and accurate performance.

One important issue for video classification remains. We can select discriminative features for a classifier for each feature set. However, in our approach, we did not deal with the problem which feature set is suitable for this discrimination task. We manually chose color histogram, face information and speaker id as representative feature sets for the anchor detector. Future work will focus on how to select representative feature sets through automatic analysis.

6. Acknowledgments

This research is partially supported by the Advanced Research and Development Activity (ARDA) under contract numbers MDA908-00-C-0037.

7. REFERENCES

[1] TRECVID, The NIST TREC Video Retrieval Evaluation, http://www-nlpir.nist.gov/projects/trecvid/.

[2] Schneiderman, H. and Kanade T. Object Detection Using the Statistics of Parts, International Journal of Computer Vision 2003.

[3] Chang, C.-C and Lin, C.-J., LIBSVM: a library for support vector machines, 2001. http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[4] Lin, W.-H., Jin, R. and Hauptmann A. Meta-classification of Multimedia Classifiers, International Workshop on Knowledge Discovery in Multimedia and Complex Data, Taipei, Taiwan, May 6, 2002

[5] Gauvain, J.L., Lamel, L. and Adda, G., The LIMSI broadcast news transcription system. Speech Communication, 37(1-2), 89-108, 2002