COMBINING TEXT AND AUDIO-VISUAL FEATURES IN VIDEO INDEXING

Shih-Fu Chang[1], R. Manmatha[2], and Tat-Seng Chua[3]

[1] Department of Electrical Engineering, Columbia University
[2] Department of Computer Science, University of Massachusetts Amherst
[3] School of Computing, National University of Singapore

ABSTRACT

We discuss the opportunities, state of the art, and open research issues in using multi-modal features in video indexing. Specifically, we focus on how imperfect text data obtained by automatic speech recognition (ASR) may be used to help solve challenging problems, such as *story segmentation, concept detection, retrieval, and topic clustering*. We review the frameworks and machine learning techniques that are used to fuse the text features with audio-visual features. Case studies showing promising performance will be described, primarily in the broadcast news video domain.

1. INTRODUCTION

Automatic indexing of video data requires solutions at multiple levels, many of which involve understanding of semantic information such as people, locations, and events. Automatically transcribed text from speech data associated with the video sequence provides a direct source for such semantic information. Though speech recognition is still imperfect in most practical situations, ASR data has been shown important for improving the overall video indexing performance.

The ASR data is most useful when the recognition targets or retrieval topics are closer to the semantic level. For example, ASR data is useful for key term/named entity extraction, story boundary detection, concept annotation, and topic change detection. On the other hand, ASR data is irrelevant for low-level tasks such as video shot segmentation, which is primarily defined by visual scene transitions and editing operations.

In this paper, we review the promising results and new directions for video indexing by combining text with features of other modalities. We focus on areas that deal with semantic-level tasks mentioned above. Most cases use the TRECVID news video benchmark [4] as the data domain to validate the performance. We expect the principles and methodologies covered here to be generalizable to other domains, although the performance will vary due to the varied levels of correlation between text and visual data in each domain.

2. FEATURE EXTRACTION

One critical issue for video indexing and related recognition problems is the selection of features that can be efficiently extracted and used to distinguish different semantic classes. On the visual side, recent works have expanded low-level features, such as color, texture, edge, and motion, to add mid-level abstractions. Among the popular ones are those related to people (face, anchor, etc), acoustics (speech, music, pitch, significant pause, etc), objects (image blobs, building, graphics, overlay text, etc), locations (indoor, studio, city, etc), genres (weather, sports, commercial, etc), and productions (camera operations, blank frames, etc) [1][2][4]. In some cases logical predicates are formed by detecting the presence and relation of the primitive features – e.g., significant pause followed by the anchor scene. Abstracting low-level features to the mid level allows for inclusion of different modalities without resulting in an excessively high dimensionality. It also allows development of statistical methods modeling the semantic relations at a higher level.

The ASR transcripts are often processed by some shallow language techniques, such as word stemming, stop word removal, rare word filtering, and part-of-speech tagging [3]. Sometimes name entities and domain-specific cue words are also extracted [2].

3. STORY SEGMENTATION

News video story segmentation offers an excellent case testifying to the power of multi-modal fusion. Video data associated with the story transition points have strong cues from audio (e.g., spectral features, music), speech (e.g., prosody, cue words), and visual (e.g., anchor, scene). In [1], it is reported that the text-based approach using ASR transcripts achieves an F_1 accuracy measure[1]

[1] $F_1 = 2/ (1/P + 1/R)$. It's a popular measure used to assess the tradeoff between precision (P) and recall (R).

of 0.62, compared to 0.69 using audio-visual features without ASR, and 0.75 using a combination of all modalities (based on the TRECVID 2003 test data set). A Maximum Entropy model was developed to select the salient features and learn the contribution weights of individual features of different modalities. Discriminative classification methods such as Support Vector Machines (SVMs) were shown to provide additional performance improvement, despite the lack of explicit models of generation likelihood. In addition, models like HMMs that capture information about the temporal transition structures and probabilities were shown to be useful for detecting story boundaries [2].

4. CONCEPT ANNOTATION

Image annotation involves using a statistical model with a training set of annotated images to automatically annotate a test set of images with keywords. The same process can be applied to annotation of video at different granularities, shots or stories. Once the annotations are produced, it is easy to retrieve images or video given text queries. Recent research efforts, such as those in the NIST TRECVID video retrieval evaluation [4] have demonstrated progress in detecting semantic concepts at an increasing scale. Examples of concepts include those related to sites (e.g., beach, indoor, city), objects (e.g., train), events (e.g., airplane taking off, people walking), or people. Multi-modality and multi-model fusion has been shown to be important in achieving good accuracy [5].

4.1. Multi-modal and Multi-model Fusion

One generic approach to concept detection is combining multiple single-feature or single-modal classifiers by ensemble fusion. Each individual classifier uses statistical models like a GMM or SVM and operates on a single feature or a small set of features. The detection scores from individual classifiers are then fused using a linear mixing function or a discriminative classifier (i.e., SVM). The weights of the linear function or the parameters of the fusing classifier are optimized through a grid search in the parameter space. A significant performance gain was demonstrated for almost all concepts when fusing classifiers of audio-visual features with those using ASR data. This can be considered as a late fusion strategy, in contrast with the early fusion approach which aggregates features of multiple modalities into a single classifier. [5].

4.2 Cross-Media Language Model

This problem of image annotation using multi-modal features lends itself easily to the application of statistical models originally proposed in the areas of human language processing. Below we survey a promising direction based on language models.

The images may be described using a visual language of visterms (analogous to words) and the keywords using English and thus the problem may be viewed as analogous to the problem of machine translation. Duygulu *et al used* IBM (translation) model 2 to solve this problem [6]. Given that word order is immaterial to keyword annotations, it turns out that IBM model 1 (which does not use word order) works much better than the other IBM translation models. Jeon *et al* [7] assumed that the problem of image annotation could be viewed as analogous to the problem of cross-lingual retrieval. They then went on to adapt relevance (based language) models for this problem and were able to show that they performed better than the translation models on the image annotation/retrieval task. They called this model the Cross-Media Relevance Model (CMRM). The CMRM computes the probability of the annotation given the image. It does this by computing a mixture over all training images. Given certain assumptions, for each training image, the computation of this probability involves the product of two components. The first is the probability of the annotation given the training image $P(w|J)$ where w is the annotation and J a training image. $P(w|J)$ is usually assumed to be a multinomial. The second the probability of each visterm of the test image given a training image $P(v\_i|J)$ where v_i is a visterm.

Another approach that has also been used previously [7] is the co-occurrence model of Mori *et al* which involves creating a co-occurrence table of visterms against words using the training set in order to annotate the test set.

The visual vocabulary of visterms may be produced in two different ways depending on whether discrete or continuous features are used. Features have typically included color and texture features or features about region extent. The discrete model involves dividing up the image into segments or regions, computing features over the regions and then creating visterms by clustering across the training images. It turns out that using a regular grid partition is better than using region segmentation [9]. This is probably more a reflection of the state of still image segmentation. Segmentation is done over a single image. On the other hand, the finer rectangular partitions allow associations to be learned over multiple images using the annotation models.

The continuous approach involves using the (continuous) features directly i.e. denoting continuous features using v_i. Blei and Jordan [8] suggested a number of different models including Correlation Dirichlet Allocation. Feng *et al* [9] proposed the continuous relevance model which is a continuous version of the relevance model and involves using a kernel density estimate to estimate $P(v\_i|J)$. The results obtained by this model are much better than other models.

5. VIDEO RETRIEVAL

5.1 Cross-Media Model for Video Retrieval

Image annotation models may also be applied to video annotation and retrieval in a number of different ways. One approach is to view each video as a succession of still frames and apply the model to each frame. For computational reasons, this is usually restricted to keyframes. The second approach assumes that videos are more than collections of still frames. One needs to model an entire clip – for example by computing features over the entire clip. Very little work has been done on the second approach partly because of the difficulty of computing appropriate features.

Based on the cross-media annotation model, we can compute the probability of an annotation concept (e.g., people, indoor) given a test image without ASR. Such probabilities can be ranked over all candidate images in the database in response to a query of concept (e.g., 'find video shots that contain people', 'find video shots of airplanes taking off'). It is observed that such a ranking process often results in good performance in terms of the precision of returned results at certain sizes (or average precision). This is an encouraging application of language models although the absolute values of the probabilities of the annotation may not be accurate.

An important consideration especially in video models is how words are modeled. Different images have different numbers of annotation words. For example, assume that image 1 has the annotation "face" while image 2 has 3 annotations "face, news-reader, desk". Using a multinomial, face has probability 1 in the first case and 1/3 in the second case. This is problematic since both images have a face. Two solutions to this problem include using a Bernoulli model for word distributions or using a multinomial after padding the annotations to fixed length. Both models give the same annotation performance while the second one does better on retrieval [9].

Annotations in training images may be produced using the closed captions or speech transcripts. Alternatively, they may also be produced by manual annotations as was done with the TRECVID data. It is important for the annotations to be visual entities since the annotation models are learning visual correspondences.

The different models have been tried on the subsets of the TRECVID dataset as well as the complete TRECVID dataset. Results show that the continuous relevance models outperform machine translation, CMRM and even the Gaussian mixture models and HMM's on this task [10]. In future work, improved results will require both better features and models.

5.2 Query Class Dependent Multi-modal Retrieval

Early results [4] showed that effective video retrieval requires the judicious use of multi-modality features to induce relevant video shots. In the news video domain, the useful features include text from ASR, audio-visual, and specialized detectors such as the video OCR, face recognizer and speaker identifier. Although text as a feature has been demonstrated to be the most reliable in retrieving a large set of relevant video shots, non-text features are found to be essential in re-ranking the text retrieval output in improving precision [12][13].

The main issue in multi-modality combination is how to fuse the features effectively. Most early systems investigated various heuristic and learning based approaches to combine features for query-independent retrieval. However, Yan *et al* [13] found that the use of learning a set of query-independent weights to combine features sometimes performed worse than a system that uses text alone, thus highlighting the difficulty of multi-modality combination. As different queries have different characteristics, it seems intuitive to explore query-dependent models for retrieval. Borrowing from the ideas used in text-based question-answering research [14], a feasible idea is to classify queries into pre-defined classes and develop fusion models by taking advantage of the prior knowledge and characteristics of each query class. Such an idea is being employed effectively in recent video retrieval systems [12][13]. Such systems essentially employ a search pipeline similar to that of text-based definition question-answering approaches. Given a query, they first perform query analysis to categorize the queries into pre-defined classes, and employ query-dependent models to fuse the multi-modal features using a linear mixture function. The query-class associated weights are trained using a learning-based approach such as the Expectation Maximization (EM) algorithm [13].

Yan *et al.* [13] considered four query classes of type named person, named object, general object and scene, and explored a 2-level hierarchical query-dependent fusion model that emphasizes text features. They tested their system on TRECVID 2003 test data and found significant improvements in retrieval performance over the use of text-only features, and the ideal query-independent model learned by assuming ground truth on the best collection available. Chua *et al* [12] further explored the use of external knowledge, specialized detectors and pseudo relevance feedback in a single-level query-dependent model with 6 query classes of type person, sports, finance, weather, disaster and general. They reported successive improvements in retrieval performance in terms of MAP (mean average Precision) from 0.071 with the use of text only feature supplemented by external knowledge from the web and WordNet, to

0.119 with the use of shot classes, video OCR, face recognizer and speaker identifier, and finally to 0.124 with the application of pseudo relevance feedback. The overall system achieves the best performance in fully automated retrieval in recent TRECVID 2004 evaluations.

6. TOPIC CLUSTERING

Videos from different times or sources can be grouped into distinct clusters, each of which is associated with a unique topic, such as 'tornado in Florida' or 'Iraqi conflict'. For concepts at such a high level, text features from ASR are without doubt very important. Satisfactory performance has been seen in automatic topic detection and tracking, a task that has been carried out in the TREC-TDT effort. An interesting question that arises is whether combination of audio-visual features with the text feature will contribute to the discovery of interesting and novel topics.

Xie *et al* proposes a layered dynamic mixture model to discover multi-modal clusters across audio, visual, and speech transcript streams [11]. To capture temporal structures, a HMM or Hierarchical HMM is first used to find clusters in audio and visual streams. Text from the speech transcript stream is clustered using latent semantic analysis (LSA), which treats stories as separate text documents. The clusters from the LSA and HMM analysis form the mid-layer tokens, over which a top-layer mixture model is developed to learn the joint probability among multi-modal tokens. Experiments with the TRECVID 2003 data set indicate such multi-modal fusion indeed results in a higher accuracy in detecting certain topics that involve strong cues from multiple modalities. The most notable among them include the topics of 'Winter Olympics', 'NBA Finals', and 'tornado in Florida'. Videos of such topics tend to have unique audio-visual features (e.g., motion, graphics, and scenes) as well as salient textual terms. An interesting direction for future research is to investigate video topic detection and tracking in the absence of ASR data as is the case for foreign news or news with poor audio quality. Techniques for multi-modal concept annotation and video retrieval, as discussed earlier, offer great potential for solving this problem.

7. CONCLUSIONS

Multi-modal fusion that combines ASR data with audio-visual features is critical for many important problems in video indexing. Recent work has shown promising results in specific areas such as story segmentation, concept detection, retrieval, and topic clustering. This field continues to present many challenges in both theory and system building. Among them, audio-visual feature selection and abstraction, especially in the temporal dimension, requires more attention. Recognition of events and activities in the video remains challenging. Better understanding and modeling of relations between concepts in the text stream and features extracted from different levels of the audio-visual streams will be essential for exploiting the full potential of multi-modal content analysis.

8. REFERENCES

[1] W. Hsu, L. Kennedy, C.-W. Huang, S.-F. Chang, C.-Y. Lin, G. Iyengar, "News Video Story Segmentation using Fusion of Multi-Level Multi-modal Features in TRECVID 2003," In IEEE ICASSP, Montreal, Canada, May 2004.

[2] T.-S. Chua, S.-F. Chang, L. Chaisorn, and W. Hsu, "Story Boundary Detection in Large Broadcast News Video Archives Techniques, Experience and Trends," ACM Multimedia, New York, Oct. 2004.

[3] L. Xie, L. Kennedy, S.-F. Chang, C.-Y. Lin, A. Divakaran, H. Sun, "Discover Meaningful Multimedia Patterns with Audio-Visual Concepts and Associated Text," IEEE Intern. Conference on Image Processing (ICIP), Singapore, Oct. 2004.

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

[5] A. Amir *et al*, "IBM Research TRECVID 2004 Video Retrieval System, NIST TRECVID 2004.

[6] P. Duygulu, K. Barnard, J. F. G de Freitas, and D. A. Forsyth, "Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary," in the 7th European Conference on Computer Vision, pages IV: 97–112, 2002.

[7] J. Jeon, V. Lavrenko and R. Manmatha, "Automatic Image Annotation and Retrieval Using Cross-Media Relevance Models," Proc. ACM SIGIR'03, pp 119-126, 2003.

[8] D. Blei and M. Jordan, "Modeling Annotated Data", Proc. ACM SIGIR'03, pp 127-134, 2003.

[9] S. L. Feng, R. Manmatha and V. Lavrenko, "Multiple Bernoulli Models for Image and Video Annotation," Proc. CVPR'04, vol, II, pp 1002-1009, 2004.

[10] G. Iyengar, P. Duygulu, S. L. Feng, P. Ircing, S. Khudanpur, D. Klakow, M. Krause, R. Manmatha, H, J. Nock, D. Petkova, B. Pytlik, P. Virga," Joint Visual-Text Modeling for Multimedia Retrieval", 2004 CLSP Workshop Report, John Hopkins University.

[11] L. Xie, L. Kennedy, S.-F. Chang, A. Divakaran, H. Sun, and C.-Y. Lin, "Layered Dynamic Mixture Model for Pattern Discovery in Asynchronous Multi-Modal Streams," in IEEE ICASSP, Philadelphia, March 2005.

[12] T.S. Chua, S.Y. Neo, K. Li, GH. Wang, R. Shi, M. Zhao, H. Xu, S. Gao and T.L. Nwe, "TRECVID 2004 Search and Feature Extraction Task by NUS PRIS". In NIST TRECVID-2004, Nov 2004.

[13] R. Yan, J. Yang, and A.G. Hauptmann, "Learning Query-Class Dependent Weights in Automatic Video Retrieval," ACM Multimedia 2004, New York, pp. 548-555.

[14] H. Yang, L. Chaisorn, Y. Zhao, S.Y. Neo and T.S. Chua, "VideoQA: Question-Answering on News Video," ACM Multimedia 2003, Berkeley, pp. 632-641.