

SEMANTIC INDEXING OF MULTIMEDIA USING AUDIO, TEXT AND VISUAL CUES

G. Iyengar, H. Nock, C. Neti, M. Franz

IBM TJ Watson Research Center, Route 134, Yorktown Heights, NY 10528. USA

ABSTRACT

In this paper we describe methods for automatic labeling of high-level semantic concepts in documentary style videos. The emphasis of this paper is on audio processing and on fusing information from multiple modalities. The work described represents initial work towards a trainable system that acquires a collection of generic “intermediate” semantic concepts across modalities (such as audio, video, text) and combines information from these modalities for automatic labeling of a “high-level” concept. Initial results suggest that multi-modal fusion achieves a 12.5% relative improvement over the best unimodal model.

1. INTRODUCTION

Large digital video libraries require tools for storing, searching and retrieving content. This paper describes methods for the automatic labeling and retrieval of high-level semantic concepts within unstructured video, with particular focus upon the integration of information from multiple modalities. The work reported represents initial results from a project which is developing a generic trainable system for classification of multimedia content based on semantics.

In common with many authors such as [1, 2, 3], we approach semantic labeling as a machine learning problem. Much of prior work focused on extracting semantic information from a single modality and joint modeling of intermediate-level concepts. The novelty of this work is in combining cues from audio, text and visual modalities for extraction of higher-level concepts (such as testimonial, rocket launch) as opposed to intermediate level concepts (such as music, outdoors, sky)¹. This paper will focus mainly upon the processing and integration of cues from audio and speech. More details of video processing and other fusion schemes may be found in [4].

The structure of the paper is as follows. Section 2 presents a detailed overview of the proposed semantic content analysis system. Sections 3 and 4 describe the processing of audio and speech streams, respectively; Section 5 describes the approaches investigated for combining multiple intermediate concept models into a single higher-level concept model. Section 6 presents experimental results. The paper ends with conclusions and future work.

2. SEMANTIC CONTENT ANALYSIS SYSTEM

Our approach to automatic semantic labeling begins by assuming the a-priori definition of a set of *intermediate-level* semantic concepts (objects, scenes and events) which is assumed to be broad enough to cover the semantic query space of interest. Such concepts can be annotated manually within a set of “training” videos.

¹Audio here refers to non-speech content of the sound-track; Text could be derived from, for example, the speech content.

Examples of intermediate concepts occurring in audio include explosions, music, speech; for video, outdoors, sky and faces. The annotated training data is used to develop explicit statistical models of the intermediate-level concepts; each such model can be used to automatically label occurrences of the corresponding concept in new videos. However, semantic concepts of interest to users often involve multiple intermediate-level concepts; the example considered in this paper is a rocket launch, which (generally) occurs outdoors and involves an explosion and often has a rocket object visible. In addition, the speech content may offer a countdown sequence. More complicated statistical models must be constructed for these *higher-level* concepts, combining (“fusing”) information from intermediate-level models and from any other usable information sources. Fusion may occur at various levels: at the level of features, across only audio-related or video-related concept models, or across concepts in multiple modalities. As with intermediate-level concepts, the resulting higher-level semantic models can then be used to label new videos. As a starting point, our unit of semantic labeling and retrieval is a camera shot. Future work will address whether this is the most effective unit for semantic concept labeling. The overall framework is illustrated in Figure 1.

There are several challenges to be overcome in such a system. Firstly, low-level features appropriate for labeling intermediate-level concepts must be identified (since different features may be appropriate for different concepts). Secondly, higher-level concepts must be linked to the presence (or absence) of appropriate intermediate-level concepts (either within a modality or across). Thirdly, a set of appropriate fusion models need to be chosen to integrate the information from multiple intermediate level concepts. In this paper, we do not address the first two and focus on the third. Future work will address techniques to automatically select relevant intermediate labels and low-level features.

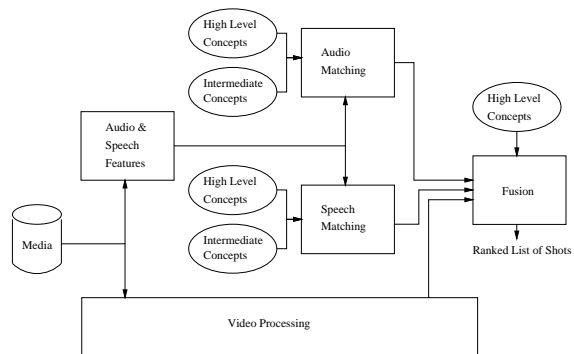


Figure 1: Diagram of Semantic Concept Analysis System

3. MODELING AUDIO-BASED INTERMEDIATE-LEVEL CONCEPTS

The scheme for modeling audio-based intermediate-level concepts, such as silence, explosions or music, begins with the annotated audio training set described earlier. Regions corresponding to each class are segmented from the audio and low-level features extracted. One obvious modeling scheme uses these features to train a Gaussian mixture model (GMM) for each concept. However, this ignores the duration properties of the audio events; use of these GMMs to label new (or even training) videos (by assigning each frame in the new data to the most likely generating concept) may yield implausibly short events. One scheme for incorporating duration modeling is as follows: a Hidden Markov Model (HMM) with some number of states is used to model each audio concept; each state in a given HMM has the same observation distribution, namely the GMM trained in the previous scheme². This can be viewed as imposition of a minimum duration constraint on the temporal extent of the intermediate-level labels.

To summarize, we used the following schemes for generating intermediate level concept scores.

- Scheme 1: The HMMs for different concepts are placed in parallel and a Viterbi decoding is used to segment the audio track into a sequence of audio-event labels. This segmentation is integrated over the duration of a shot which gives us the fraction of a shot for which the particular concept was present.
- Scheme 2: A more refined method to estimate the fractional presence of the different intermediate-level concepts in a shot is to use the HMMs to generate an N-best list at each audio frame and then average these scores over the duration of the shot³.
- Scheme 3: We notice that there are variations in the absolute values of these scores due to variations in the shot lengths, thresholds chosen for generating the N-best list etc. To counter these variations, we *normalize* these scores by dividing each concept score with the sum of all the concept scores in a particular shot.

We note that Schemes 2 and 3 outperform Scheme 1 and therefore restrict further discussion to the latter two.

4. USING SPEECH FOR RETRIEVAL OF HIGHER-LEVEL CONCEPTS

In addition to the audio-based intermediate-level concepts, we use textual cues derived from the speech content in the video. This can be either from manual transcriptions such as close-captioning or produced using an automatic speech recognition (“ASR”) system on the speech content. Given corpus transcriptions of either type, the transcriptions must be split into documents. Documents are defined here in two ways: the words corresponding to a shot, or the words in a shot plus its left and right neighbors. (The latter reflects a belief that in highly edited videos, speech cues may occur not just within the unit of the (potentially short) shot, but also in surrounding shots; “surrounding shots” might profitably be defined as “shots in the same scene as the shot of interest”, but there is no scene detection in the current system. Use of left and right

²It is closely related to the speech vs. non-speech segmentation scheme of IBMSpine2, see Saon et al., ICASSP02

³N-best lists can be generated in multiple ways. We specifically used the scheme suggested in Zweig and Padmanabhan, ICSLP 2000.

neighboring shots as an alternative follows [5].) The word time marks necessary to determine the mapping from word tokens to shots can be obtained using either a forced alignment with an ASR system (for ground truth transcriptions) or directly from the ASR output (for the case of automatically produced transcriptions). The words in each document are then tagged with part-of-speech (eg. noun phrase), which enables morphological decomposition to reduce each word to its morph. Finally, stop words are removed using a standard stop-words list.

Our system for retrieving a particular semantic concept using speech transcriptions alone assumes the a-priori definition of a set of pertinent query terms. The set might be derived from human knowledge, Word Net [6] or from words co-occurring frequently with that concept in the annotated training data. Tagging, morphologically analyzing and applying the stop list to this set of words (in the same way as was applied to the database documents) yields a set of query terms Q . Database documents are ranked against Q by their *OKAPI* score [7].

5. FUSION OF CUES FROM AUDIO, TEXT AND VISUAL MODELS

Processing of the individual modalities yields a set of scores for each shot in the corpus from Audio, Text and Visual models. The scores derived from the processing of individual modalities can then be combined to produce a final ranking of documents. We note here that the intermediate visual models include fire, sky, outdoors, face, and rocket vehicle and these models were trained using Support Vector Machines.

For fusing the scores, we used the following approach. The scores from all the intermediate models were combined to form a vector whose dimensionality is the number of intermediate concepts. A Support Vector Machine (SVM) is then trained in this space with positive and negative examples of the high-level concept (in our case, the rocket-launch event). This is illustrated in Figure 2. We note here that we also experimented with Bayesian Networks (BN) for fusion between different modalities. The performance of the BN framework for fusion is not as good as the SVM model and we restrict further discussion to the SVM model.

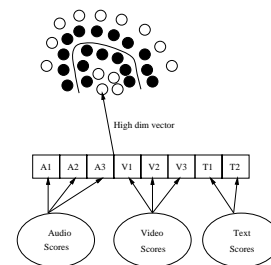


Figure 2: Combining Information From Multiple Intermediate Concepts using Support Vector Machines

6. EXPERIMENTAL RESULTS

The following results demonstrate the feasibility of our semantic labeling framework for a single high-level concept: “rocket launch”. All labeling (or retrieval) is applied at the level of camera shots.

6.1. Corpus and Shot Segmentation

Our experiments use subsets of the NIST Video TREC 2001 corpus [8], which comprises production videos derived from sources such as NASA, and OpenVideo consortium. Shot segmentation was performed using the *IBM CueVideo* toolkit [9, 10]. The corpus supplies evidence to support our hypothesis that integration of cues from multiple modalities is necessary to achieve good labeling or retrieval performance. Of the 78 manually annotated rocket launch shots, only 51 contain speech and only a subset of those contain rocket-launch related words. (The subset when errorfull ASR is in use is potentially even smaller.) For the same 78 rocket-launch shots, the most pertinent audio cues are music and explosions, found in 84% and 60% of manually labeled audio samples respectively. In the visual side, the rocket shots are from a variety of poses and in many cases the rocket exhaust completely occludes the rocket object. Therefore, it seems unlikely that any single audio, speech or visual cue could retrieve all relevant examples.

6.2. Audio-only Retrieval Results

The current set of audio-based intermediate-level semantic concepts comprises six events including explosion, music, speech, noise, silence, and nonsense. Examples of these events have been manually annotated in the sound-track of 13 video clips from TREC 2001. Since data labeled with these events is limited, a cross-validation or leaving-one-out strategy is adopted in these experiments: models are trained on all-but-one video and tested on the held-out video. The results presented are the average over all such combinations. The intermediate-level labels thus generated are combined for the high-level “rocket launch” event retrieval.

We used 24-dim Mel-Frequency Cepstral Coefficients, common in ASR systems, as our low-level features. Features used by other authors include centroid frequency, pitch etc [1, 11, 12, 13].

In the first experiment, we study the effect of using a HMM for duration modeling of a single intermediate concept (Explosion). In the second experiment, we look at the effect of different audio-only fusion strategies. We note here that Scheme 3 in section 3 can be viewed as *implicit* fusion where the score for a concept is now based on all other concept scores in a shot. For explicit fusion, we take the scores (from Scheme 2) of explosion, music, speech and speech-music and combine them using a Bayesian Network. In implicit fusion, we rank the shots based on the normalized (Scheme 3) score of the explosion cue.

Figure 3, part (a), compares the retrieval of the intermediate-level explosion concept with HMM and GMM scores, respectively. Notice that the HMM model has significantly higher precision for all recall values compared to the GMM model. Figure 3, part (b), compares rocket-launch concept retrieval when using the implicit and explicit schemes to combine multiple intermediate-level audio concept scores. Notice that the implicit fusion scheme has a significantly higher precision for all recall values.

6.3. Speech-only Retrieval Results

The speech-only retrieval experiments use a subset of 13 video clips from TREC 2001. Two retrieval scenarios are investigated: retrieval using ground-truth transcriptions and retrieval using transcriptions produced using ASR. The ground truth transcriptions comprise manually-produced transcriptions; words were time aligned to shots using the IBM HUB4 (Broadcast News) ASR system [14]. The speech recognition transcriptions were produced using the

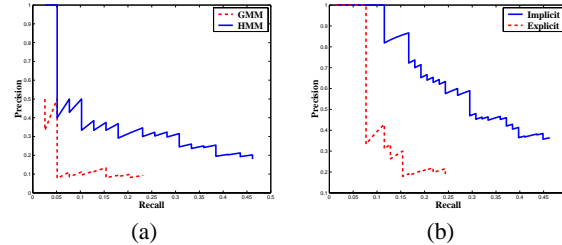


Figure 3: (a) Effect of Duration Modeling and (b) Implicit vs Explicit Fusion

same system. Prior to generating the ASR, the audio data was preprocessed using an automatic speech/non-speech segmenter⁴. The frame-level accuracy of the segmenter is 77% (speech 88%, non-speech 59%). Very short segments are then merged. The ASR error rate over the manually transcribed 13 video retrieval subset is currently 47.6%. Documents derived from these transcriptions comprise words corresponding to single shots or to shots plus their left and right neighbors.

Two query term sets Q pertinent to rocket launches were used: the first **training-set based query** comprises query terms selected from amongst words frequent in rocket launch shots (*engines, flight, lift, off, NASA, five, four, three, two, one, shuttle, space*) and the second **human-knowledge based query** is obtained by asking users unfamiliar with the TREC corpus for words expected to be pertinent for the rocket launch event. (*NASA, ariane, rocket, launch, space, agency, nasda, satellite, spacecraft, space, shuttle, mission*).

Figure 4 shows the performance of the text-based retrieval for the rocket launch query. It is clear that improving the ASR can improve the retrieval performance. However, clearly more work is needed in query formulation and expansion.

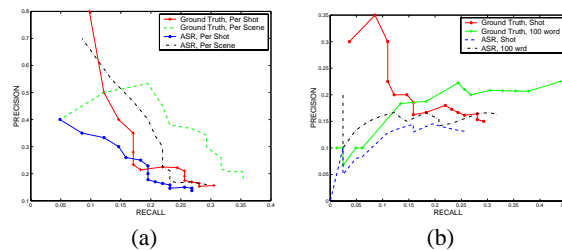


Figure 4: (a) Results: Training-set based Query. (b) Results: Human-knowledge based Query

6.4. Fusion of Multiple Modalities for Retrieval

For fusion using Support Vector Machines, we used 10 intermediate and high-level models (Audio: explosion, music, speech, speech-music; Video: rocket, outdoors, sky, face, fire-smoke; Text: rocket launch) with a radial basis function (RBF) kernel⁵. The results are reported using leave-one-video-out cross validation.

Figure 5 illustrates the performance of retrieval using audio model, video model and the joint SVM fusion model. For audio, we used the implicit fusion model. For video, we used the rocket object

⁴Using a scheme similar to IBMSpine2 system

⁵using SVMLight <http://svmlight.joachims.org/>

| Technique | Retrieval FOM |
|------------------------------|---------------|
| Best uni-modal (audio) | 0.56 |
| Best visual | 0.39 |
| Text (human-knowledge based) | 0.14 |
| Joint (audio+text+visual) | 0.63 |

Table 1: Summary of retrieval experiments using individual modalities and the joint SVM model

SVM model [4]⁶. The text-only model is not shown in the plot because of its poor performance. To enable a quick comparison of the precision-recall figures, we calculate an overall figure-of-merit (FOM) which is the average precision over the top 100 retrieved documents. Table 6.4 presents the FOM results for the various retrieval methods. We note that the multi-modal fusion model performs better than the individual modalities. It appears that for this particular high-level concept, the audio cue is fairly strong. When we test on multiple high-level concepts (such as say rocket launch and airplane takeoff, panel discussion and newscast), the power of multiple modalities might be more apparent as individual modalities might suffer from confusions.

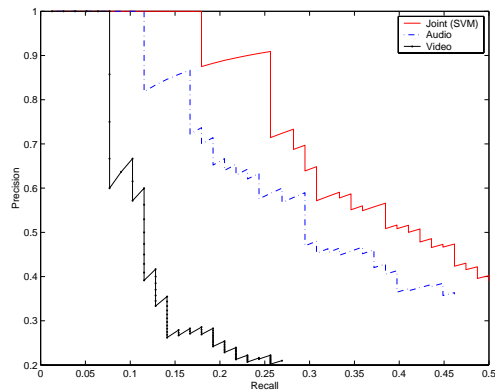


Figure 5: Fusion of Audio, Visual and Text models using SVM for rocket launch retrieval

7. CONCLUSIONS

Feasibility of the fusion approach for retrieving high-level concepts was demonstrated for the semantic concept “rocket launch”, with particular emphasis on the integration of cues from audio, speech and video. The experimental results, whilst preliminary, suffice to show that information from multiple modalities (visual, audio, speech and potentially video-text) can be successfully integrated to improve semantic labeling performance over that achieved by any single modality.

There is considerable scope for improving the schemes described for semantic labeling using audio and text information. Future research directions include the utility of multi-modal fusion at earlier stages (using, for example, coupled HMMs or other dynamic Bayesian networks), the appropriateness of shot-level rather than frame-level (or other) labeling schemes. Schemes must also be identified for automatically determining the low-level features (from a pre-defined set of possibilities) which are most appropriate for labeling intermediate-level concepts, and for determining intermediate level concepts (amongst the predefined set of possibilities)

⁶This was trained and tested on human-labeled bounding boxes.

which are related to higher-level semantic concepts.

8. ACKNOWLEDGMENTS

Brian Kingsbury, George Saon, Satya Dharanipragada, Benoit Maisson and other members of the HLT group and M. Naphade and C-Y. Lin for their video features and models.

9. REFERENCES

- [1] Michael A. Casey, “Reduced-rank spectra and minimum-entropy priors as consistent and reliable cues for generalized sound recognition,” in *Proceedings of Eurospeech*, 2001.
- [2] Milind R. Naphade and Thomas S. Huang, “A probabilistic framework for semantic video indexing, filtering and retrieval,” *IEEE Transactions on Multimedia, special issue on Multimedia over IP*, vol. 3, no. 1, pp. 141–151, Mar. 2001.
- [3] G. Iyengar and A. B. Lippman, “Models for automatic classification of video sequences,” in *Storage and Retrieval from Image and Video Databases*. Jan 1998, vol. VI, SPIE.
- [4] M. R. Naphade, C-Y. Lin, and J. R. Smith, “Learning semantic multimedia representations from a small set of examples,” in *Submitted to IEEE International Conference on Multimedia and Expo*, 2002.
- [5] The Lowlands Team, “Lazy users and automatic video retrieval tools in (the) lowlands,” in *Proceedings of the Tenth Text REtrieval Conference (TREC-10)*. 2001, NIST Special Publication.
- [6] Christiane Fellbaum, *WordNet: An Electronic Lexical Database*, MIT Press, Cambridge, MA, USA, 1998.
- [7] M. Franz and S. Roukos, “Trec-6 ad-hoc retrieval,” in *Proceedings of the Sixth Text REtrieval Conference (TREC-6)*. 1998, pp. 511–516, NIST Special Publication 500-240.
- [8] NIST, “The Tenth Text REtrieval Conference TREC10,” in http://trec.nist.gov/pubs/trec10/t10_proceedings.html, 2001.
- [9] IBM Almaden Research Center, “The IBM cuevideo project,” in <http://www.almaden.ibm.com/cs/cuevideo/index.html>, 1997.
- [10] J. R. Smith, S. Srinivasan, A. Amir, S. Basu, G. Iyengar, C. Lin, M. Naphade, D. Poncelon, and B. Tseng, “Integrating features, models, and semantics for trec video retrieval,” in *Proceedings of the Tenth Text REtrieval Conference (TREC-10)*. 2001, NIST Special Publication.
- [11] Z. Liu, Y. Wang, and T. Chen, “Audio feature extraction and analysis for scene segmentation and classification,” in *Journal of VLSI Signal Processing System*, June 1998.
- [12] T. Zhang and C. Kuo, “Content-based classification and retrieval of audio,” in *SPIE’s 43rd Annual Meeting - Conference on Advanced Signal Processing Algorithms, Architectures, and Implementations VIII, San Diego*, July 1998.
- [13] E. Scheirer and M. Slaney, “Construction and evaluation of a robust multifeature speech/music discriminator,” in *Proc. ICASSP ’97*, Munich, Germany, 1997, pp. 1331–1334.
- [14] R. Bakis, S. Schen, P. Gopalakrishnan, R. Gopinath, S. Maes, and L. Polymenakos, “Transcription of broadcast news system: Robustness issues and adaptation techniques,” in *Proc. ICASSP ’97*, Munich, Germany, 1997, pp. 711–714.