RC-22873 (W0308-053) August 14, 2003
Computer Science

IBM Research Report

Discriminative Model Fusion for Semantic Concept Detection and Annotation in Video

Giridharan Iyengar, Harriet J. Nock, Chalapathy Neti
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Research Division
Almaden - Austin - Beijing - Delhi - Haifa - India - T. J. Watson - Tokyo - Zurich

Discriminative Model Fusion for Semantic Concept Detection and Annotation in Video

G. Iyengar H.J. Nock C. Neti
IBM TJ Watson Research
Center, NY, USA.
{giyengar,hnock,cneti}@us.ibm.com

ABSTRACT

In this paper we describe a general information fusion algorithm that can be used to incorporate multimodal cues in building user-defined semantic concept models. We compare this technique with a Bayesian Network-based approach on a semantic concept detection task. Results indicate that this technique yields superior performance. We demonstrate this approach further by building classifiers of arbitrary concepts in a score space defined by a pre-deployed set of multimodal concepts. Results show annotation for user-defined concepts both in and outside the pre-deployed set is competitive with our best video-only models on the TREC Video 2002 corpus.

Categories and Subject Descriptors

H.3.3 [**Information Storage And Retrieval**]: Information Search and Retrieval—*retrieval models*

General Terms

Algorithms

Keywords

ACM proceedings, semantic concept detection, digital video annotation and indexing

1. INTRODUCTION

Large digital video libraries require tools for representing, searching and retrieving content. One possibility is the query-by-example (QBE) approach, in which users provide (usually visual) examples of the content that they seek. However, such schemes have some obvious limitations and since most users wish to search in terms of semantic concepts rather than by visual content [7], work in the video retrieval area has begun to shift from QBE to query-by-keyword (QBK) approaches which allow users to search by specifying their query in terms of semantic concepts. Query using keywords representing semantic concepts has motivated recent research in semantic media indexing [3, 10]. Recent attempts to introduce semantics in the structuring and classification of videos include [4, 9]. In these works, the emphasis has been on the extraction of semantics from individual modalities and in some instances, using audio and visual modalities. Our research work combines content analysis with information retrieval in a unified setting for semantic labeling of multimedia content using audio, visual and textual modalities. In this paper we present a novel information fusion algorithm that leverages a pre-deployed collection of semantic concept detectors (unimodal and multimodal) to further improve their performance or to detect novel semantic concepts using this inventory of concept detectors.

The performance of semantic modeling systems can be measured along (at least) three distinct dimensions that we term as: **accuracy**, **acquisition** and **coverage**. Along the accuracy dimension, the performance of individual concept detectors (and systems as a whole) are compared using standard pattern recognition metrics such as Precision-Recall, Mean Average Precision[1]. We define coverage in terms of the "number" of distinct concept models that a particular system can reliably model. For specific domains it may be possible to define coverage more formally in terms of the space of concepts occuring in that domain. The ease-of-acquisition dimension measures the system's capacity to learn new concepts with as little manual intervention as needed. In this paper, we conduct experiments to evaluate our system's performance along the coverage and accuracy dimensions.

The rest of the paper is organized as follows. In Section 2 we outline the architecture of our trainable concept annotation system. In Section 3 we present our novel discriminative model fusion approach for combining cues from multiple modalities to both improve existing concept detectors and develop novel concept detectors. In Section 4 we evaluate the performance of this approach on the TREC Video Track 2001 and 2002 corpora. The paper ends with conclusions and discussion.

2. SEMANTIC INDEXING OF MULTIMEDIA USING AUDIO, TEXTUAL AND VISUAL CUES

We assume the user defines a set ("*lexicon*") of semantic

[1]Mean Average Precision is a system-wide number used by NIST to compare retrieval systems. See http://trec.nist.gov/pubs/trec10/t10_proceedings.html for its definition.

MM'03, November 2–8, Berkeley, California, USA.

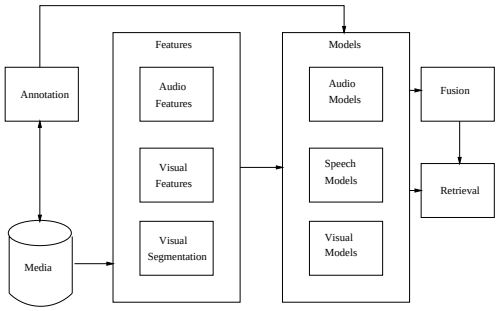

Figure 1: Automatic Annotation System Overview

concepts (objects, scenes and events) that covers their semantic query space of interest. We further assume existence of shot-level manually annotated examples for a small set of "training" videos[2]. Figure 1 illustrates the general architecture, which presents several research challenges. These include the need to develop algorithms which extract sufficiently informative low-level feature representations from manually annotated examples and to formulate a generic framework for constructing semantic concept models using the extracted features. This paper describes our approach to the generic semantic modeling framework.

3. DISCRIMINATIVE MODEL FUSION (DMF) FOR INFORMATION FUSION FROM MULTIPLE MODALITIES

Assume that a pre-defined set of "anchor" or "basis" detectors have been defined, trained and deployed in the system. See for instance, our related work in semantic concept modeling [2]. Each of these anchor model detectors can be used to score every shot in a new video. We note that these scores can be used for retrieval of these concepts from previously unseen video [2]. The scores can be likelihood ratios, log-likelihoods, SVM classification scores, result of the OKAPI formula etc, depending on the modeling approach applied for the particular semantic concept.

Once system installation begins, the user defines a lexicon and annotates examples; lexicon entries are arbitrary and may or may not correspond to concepts in the basis model set. The manually annotated examples for each concept are supplied to the system and models constructed. Each new single concept model is then constructed as follows. Each shot in the training set is scored using the predeployed basis models, giving a vector of model scores for the shot. Then, a classifier is trained to map from these vectors in "model score space" to presence or absence of the concept in a shot. See Figure 2 for an illustration. For an earlier description of this experiment, please refer to our paper [5]. This paper reports improved results obtained more recently.

We study two cases empirically: (a) target semantic concept is already a member of the basis set, (b) target semantic concept is not in the basis set. In the first case, we measure the system's accuracy. The second case is an illustration of the system's coverage capabilities. These cases are examined in the next section. Many further questions arise. What is an effective set of basis models and should more

[2]Tools for manual concept annotation in speech, non-speech audio and (or) video images are now available eg. IBM VideoAnnex, http://www.alphaworks.ibm.com/tech/multimodalannotation.

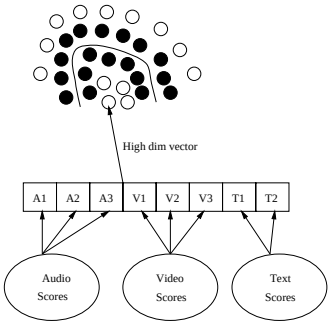

Figure 2: Illustration of Score Space Classification Approach

than one basis model be included per core concept? How large must the basis model set be to give adequate coverage of a user's semantic query space of interest? How does per-concept classification performance vary as basis set size increases? What is an appropriate choice of model score space classifier? How much user-supplied data is necessary? These questions are deferred to future work.

In this work we use a support vector machine (SVM) [8] to train a discriminative classifier in the model vector space. We can view the anchor concept classifiers as non-linear functions that take points in $\Re^n$ and map them into a scalar. That is, $C(\mathbf{x}) : \Re^n \mapsto \Re$ where $\mathbf{x}$ is an n-dimensional feature vector and C is a classifier that operates on this feature vector. We hypothesize that points that are near in the feature space produce similar scores when operated on by these classifiers. Now, if you consider a cluster in the feature space, this maps into a 1-dimensional cluster of scores for any given classifier. Extending this reasoning to multiple classifiers, we can view the SVM for fusion as operating in this new *feature* space (of classifier scores) and finding a decision boundary. In a typical situation with image features, the input feature space can be fairly large compared to the number of classifiers and here we expect the resulting dimensionality reduction to be useful.

4. EXPERIMENTS WITH DMF

We now report on two experiments. In the first experiment, we evaluate the novel DMF information fusion approach for detection of a single multimodal concept. In the second set of experiments we compare the DMF approach with direct modeling of semantic concepts to evaluate it along the accuracy and coverage dimensions.

4.1 Detection of a single multimodal concept

Corpus and Evaluation Metrics: For the first experiment reported in this paper, we use a subset of the NIST Video TREC 2001 corpus, which comprises production videos derived from sources such as NASA and Open-Video Consortium. Some of the clips contain footage of NASA activities including the space program.

We measure concept detection performance using precision-recall curves. Precision is defined as Number of relevant documents (shots)/Total retrieved documents and Recall is defined as Number of relevant documents/Total number of relevant documents in the database. In addition, an overall figure-of-merit of retrieval effectiveness is used to summarize performance. We choose the NIST Average Precision (AP) as our figure-of-merit.

Technique	AP
Best audio (explosion)	0.56
Best visual (rocket object)	0.39
DMF (audio,text,visual)	0.63
NaiveBayes (best audio + best visual)	0.59

Table 1: Average Precision results for rocket-launch detection using DMF, NaiveBayes and the uni-modal tecniques.

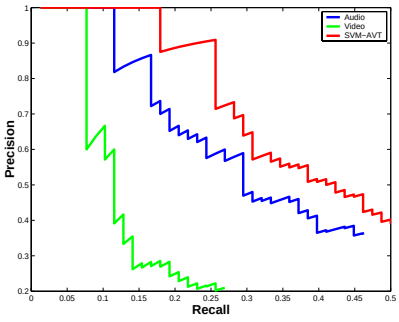

Figure 3: Fusion of Audio, Text and Visual models using the DMF model for rocket launch detection.

The target multimodal concept we choose to model is a "rocket-launch", given its strong cues in both audio and visual modalities. We first compare the performance of the DMF approach with the best audio and video detectors for rocket-launches. We note here that the best audio model for detecting rocket-launches is the "explosion" (rocket engine) sound and the best video model is the "rocket-object" model. For deteails on these audio and video models, please refer to our paper [1]. For training the Support Vector Machine in the DMF framework, we took scores from 9 semantic models (Audio: explosion, music, speech, speech-music; Video: rocket, outdoors, sky, fire-smoke; Text: rocket launch), concatenating them into a 9-dimensional feature vector. Figure 3 shows the precision-recall curve for this experiment. Figure 4 shows the screen shot of rocket-launch concept detection using the DMF model. Note that the DMF approach improves over the best uni-modal detection by 12.5% relative. Also, 19 of the top 20 retrieved shots contain rocket-launches, indicating high precision in retrieval.

We then compare the DMF approach with a Naive Bayes model integrating the Audio and Video scores[3]. Results in Table 1 shows the average precision for the two fusion approaches in addition to the 2 uni-modal techniques. The DMF approach outperforms the Naive Bayes model.

4.2 Evaluation of Concept Accuracy and Coverage of the DMF framework

Corpus: In this experiment, we use the TREC Video Track 2002 corpus, comprising 70 hours of MPEG video[4]. We use 25 hours for training basic low-level feature models ("FeatureTrain") and 5 hours for optimising parameters of individual models ("FeatureValidate") and for developing the DMF models. Final performance is evaluated on a distinct 5 hour test set ("SearchTestSubset"). The reported

[3]We implemented a variety of Bayesian networks, some including text scores and found the above Naive Bayes implementation to have the best performance.

[4]Average video length is 10 minutes.

Figure 4: The top 20 video shots of rocket launch/take-off retrieved using multi-modal detection based on the DMF model

results are based on the performance of these detectors on SearchTestSubset. We developed a variety of Text, Video and Audio models to form our basis set of detectors. These models are enumerated below.

Video Models: Concepts modeled include the six Video TREC 2002 benchmark visual concepts ("indoors", "outdoors", "face", "people", "cityscape", "landscape") plus 34 additional concepts including "sky", "transportation" and "beach". For each concept, multiple classifiers are built and their per-shot scores linearly interpolated [2].

Non-Speech Audio Models: Concepts modelled include Hidden Markov Models of "Speech" and "Instrumental Sounds" [2].

Text Models: Automatic speech recognition gives "FeatureTrain" transcripts, which are analyzed to extract all words occurring in or close to positive exemplars of each concept in the TREC 2002 visual set (above). Manual list refinement gives a set of pertinent query terms for each concept. Test set annotations are performed by first indexing the speech transcript using an OKAPI-based [6] spoken document retrieval system and then (for each concept) querying using the corresponding pertinent term set.

Basis Model Sets: For the accuracy task, we use the 40 Video, 2 Non-speech Audio and 6 Speech Models just described. For the coverage task we remove the six Video TREC 2002 visual concepts (above) from this set.

Accuracy Task and Results: The first experiment considers the case where a user concept-of-interest falls in the pre-deployed basis set. Table 2 shows per-concept Average Precision (AP) and overall Mean AP (MAP) results on the six TREC 2002 visual concepts. It is interesting to note that speech-only results for "outdoors", "face","cityscape" are comparable to best single video-only model performance. More importantly, the DMF MAP with a multimodal basis set improves performance by 19% over our pre-deployed video-only detectors; for comparison, the DMF approach using a basis of only video score vectors improves MAP by 8%. The model-score space approach does not hurt (in fact, helps) the accuracy of the underlying classifier.

Coverage Task and Results: The second experiment

Concept	Best Video Detector	Speech Detector	Video+ Speech DMF	Video DMF
Outdoors	.59	.58	.58	.58
Indoors	.12	.07	.25	.18
Face	.17	.15	.21	.18
People	.18	.18	.26	.25
Cityscape	.31	.34	.35	.30
Landscape	.19	.14	.18	.18
MAP	.26	.24	.31	.28

Table 2: Concept Accuracy AP & Overall MAP on SearchTestSubset

considers the case where a user concept-of-interest falls outside the pre-deployed basis set. We repeat the previous experiment using the reduced basis discussed above. Table 3 shows that DMF improves MAP performance by 23% compared with the best video-only TREC 2002 detectors. This indicates that the DMF approach has constructed useful models for new concepts from a pre-deployed basis. This is a very encouraging result. This enables a clear, robust approach for incorporating new concept detectors on a deployed system. Thus, the coverage of an annotation and retrieval system can be easily enhanced to incorporate concepts of users' interest.

Concept	Best Video Detector	Video+Speech DMF
Outdoors	.59	0.64
Indoors	.12	0.27
Face	.17	0.15
People	.18	0.28
Cityscape	.31	0.35
Landscape	.19	0.18
MAP	.26	0.32

Table 3: Concept Coverage AP & Overall MAP on SearchTestSubset

The DMF approach incorporates inter-model context information together (from the concatenation of model scores in forming the model-score-space vector) with individual model performance information. We hypothesize that the gains observed by using the DMF approach results from the discriminative classifier learning both these additional pieces of information. The open questions include the effect of the number of basis concepts and the choice of basis concepts on both the accuracy and coverage performance of the DMF framework. Whilst we have used SVMs as our DMF classifier, other choices are feasible. In our experiments, the basis classifiers were trained on the FeatureTrain dataset and optimized on the much smaller FeatureValidate dataset, whereas the DMF classifiers were trained on FeatureValidate. The effect of the training set size on DMF classifer needs to be quantified.

5. CONCLUSIONS

In this paper we describe a general approach for information fusion from multiple modalities in a system for automatically annotating arbitrary semantic concepts in video. In addition, this approach works well compared with proven techniques such as Naive Bayes. In a system with predeployed set of optimized basis models, we demonstrated that this novel approach provides a further gain over these models. Notably, in the Video TREC 2002 corpus we obtain a 19% improvement in MAP score. In addition, we demonstrated that this approach can be used for building arbitrary models from a set of basis concept models thereby providing a powerful general approach for video retrieval system development. On a limited test with 6 unseen concepts, the improvements in MAP was 23% over directly modeling these concepts.

We hypothesize that the exploitation of correlation (positive and negative) between the basis and target concepts is key to the performance gains offered by the DMF framework. The **accuracy** and **coverage** improvements of this approach offer a robust, consistent framework for enhancing the performance of any video retrieval system. There are many open issues, including the selection of an optimal basis, the capacity of the system to acquire novel concepts (given a set of fixed pre-deployed basis concept detectors) and training set size which need to be studied further.

6. ACKNOWLEDGEMENTS

Martin Franz, John R. Smith of IBM TJ Watson Research Center.

7. REFERENCES

[1] W. Adams, G. Iyengar, C.-Y. Lin, et. al Semantic Indexing of Multimedia Content Using Visual, Audio and Text Cues. *Eurasip JASP.*, 2:170–185, 2003.

[2] W. H. Adams, A. Amir, C. Dorai, et. al Ibm research TREC-2002 video retrieval system. In E. M. Voorhees and D. K. Harman, editors, *Proc. TREC-11*, Gaithersburg, MD, 2003. NIST.

[3] S. F. Chang, W. Chen, and H. Sundaram. Semantic visual templates - linking features to semantics. In *Proc. ICIP*, volume 3, pages 531–535, Chicago, IL, October 1998. IEEE.

[4] G. Iyengar and A. B. Lippman. Models for automatic classification of video sequences. In *Storage and Retrieval from Image and Video Databases*, volume VI. SPIE, Jan 1998.

[5] H. J. Nock, W. H. Adams, and G. Iyengar et. al. User-trainable video annotation using multimodal cues. In *Proc. SIGIR*, Toronto, Canada, July 2003. ACM.

[6] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. In *Proc. TREC-3*, pages 109–126. NIST Special Publication 500-226, 1995.

[7] J. R. Smith and S.-F. Chang. Visualseek: a fully automated content-based query system. In *Proc. fourth intl. conf. multimedia*, pages 87–92. ACM, May 1996.

[8] V. Vapnik. *The Nature of Statistical Learning Theory.* Springer, New York, USA, 1995.

[9] N. Vasconcelos and A. Lippman. Bayesian modeling of video editing and structure: Semantic features for video summarization and browsing. In *Proc. ICIP*, volume 2, pages 550–555, Chicago IL, October 1998. IEEE.

[10] T. Zhang and C. Kuo. An integrated approach to multimodal media content analysis. In *Storage and Retrieval from Image and Video Databases*, volume 3972, pages 506–517, San Jose, CA, January 2000. SPIE.