

User-trainable Video Annotation Using Multimodal Cues

H.J. Nock

W. Adams
IBM TJ Watson Research
Center, NY, USA.

G. Iyengar

{hnock,giyengar}@us.ibm.com

ABSTRACT

This paper describes progress towards a general framework for incorporating multimodal cues into a trainable system for automatically annotating user-defined semantic concepts in broadcast video. Models of arbitrary concepts are constructed by building classifiers in a score space defined by a pre-deployed set of multimodal models. 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. An interesting side result shows speech-only models give performance comparable to our best video-only models for detecting visual concepts such as “outdoors”, “face” and “cityscape”.

Categories and Subject Descriptors

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

General Terms

Algorithms

Keywords

ACM proceedings, digital video annotation and indexing

1. INTRODUCTION

Indexing and retrieval tools for digital media are an active research area at present. This paper considers the problem of automatically indexing multimedia data in terms of semantic concepts (ie. objects, events, scenes). Most current automatic semantic concept annotation systems use concept-specific algorithms (eg. face detectors [6]). However, since both content and size of the set of relevant concepts will be highly user dependent, it may not be realistic to use concept-specific techniques if systems are to be widely deployed. This motivates our research, which is developing a generic framework for incorporating multimodal cues into a system for automatically labelling arbitrary semantic concepts in broadcast video¹. Paper organisation is as follows. Section 2 outlines the high-level architecture of our trainable concept annotation system. Section 3 discusses one system component, a generic framework for integrating multimodal cues into models of arbitrary semantic concepts. Section 4 evaluates performance on the TREC Video Track 2002 corpus. The paper ends with conclusions and discussion.

2. SYSTEM OVERVIEW

We expect the user to define a set (“*lexicon*”) of semantic concepts (objects, scenes and events) that covers their

¹Multimodal cues include those in video images, speech, non-speech audio and closed captioning if available.

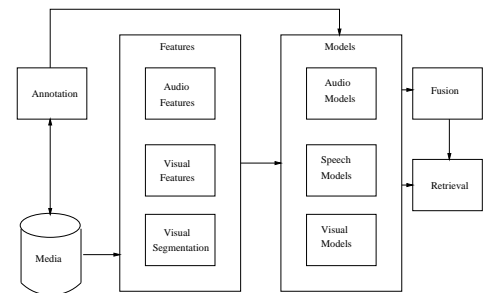


Figure 1: Automatic Annotation System Overview

semantic query space of interest and to supply shot-level manually annotated examples for a small set of “training” videos². The annotated examples for each concept are then fed into a generic framework for training statistical model(s), such as in Section 3. Once trained, the new semantic concept model(s) can be used to automatically annotate new videos at the level of shots. 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 first steps towards a generic modelling framework.

3. GENERIC CONCEPT ANNOTATION

Assume that the system is deployed with a pre-defined set of “anchor” or “basis” models, which represent some core semantic concepts such as faces, speech, indoors and cityscape. 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. A single concept model is constructed as follows. Each shot in the training set is scored using the basis models, giving per-shot vectors of model scores. Then, a classifier is trained to map from these vectors in “model score space” to presence or absence of the concept in a shot. Figure 2 illustrates use of the resulting classifier in annotating novel data, for the case of three “basis models” (“face”, “engine_noise”, “outdoors”). There will be two cases to study empirically: (a) target semantic concept is already a member of the basis set³; (b)

²Tools for manual concept annotation in speech, non-speech audio and (or) video images are now available eg. [4, 1]. Our real-time annotation factor is video-dependent, but averages 10-15 × (audio) and 1-5 × (keyframe: low end is for global annotation and high end is for marking relevant regions).

³Eg. for Figure 2, the concept to detect is one of “face”,

target semantic concept is not in the basis set. These cases are examined in the next section. Many further questions arise. What is an effective set of basis models and should more 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.

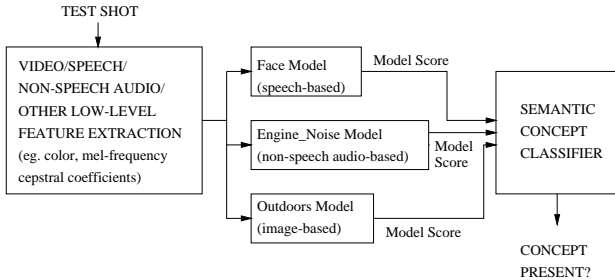


Figure 2: Generic Concept Modelling Framework

4. EXPERIMENTAL EVALUATION

Corpus: We use the TREC Video Track 2002 corpus, comprising 70 hours of MPEG video⁴. We use 25 hours for training basic low-level feature models (“FeatureTrain”) and 5 hours for optimising parameters of combined models (“FeatureValidate”); a distinct 5 hours represents (user-supplied) development data (“FeatureTest”). Final performance is evaluated on another distinct 5 hour test set (“SearchTestSubset”).

Video Models: Concepts modelled 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, multiple models are built and their per-shot scores linearly interpolated [2].

Non-Speech Audio Models: Hidden Markov Models of “Speech” and “Instrumental Sounds” are used [2].

Speech Models: Automatic speech recognition gives “FeatureTrain” transcripts, which are analysed 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 [5] spoken document retrieval system and then (for each concept) querying using the corresponding pertinent term set.

Basis Model Sets: Task “IN-BASIS” uses the 40 Video, 2 Non-speech Audio and 6 Speech Models just described. Task “OUT-OF-BASIS” removes the six Video TREC 2002 visual concepts (above) from this set.

Score Vector Classification Implementation: Support Vector Machines (SVMs) are used since early work found they outperform Bayesian networks on a similar task [3].

Task “IN-BASIS” and Results: The first experiment considers the case where a user concept-of-interest falls in the pre-deployed basis set. Table 1 shows per-concept Average Precision (AP)⁵ and overall Mean AP (MAP) results on the six TREC 2002 visual concepts. Note in passing that speech-only results for “outdoors”, “face”, “cityscape” are comparable to best single video-only model performance, an interesting side-result. More importantly, score vector MAP

⁴“engine_noise” or “outdoors”.

⁴Average video length is 10 minutes.

⁵Video TREC 2002 definition.

Concept	Best Video Detector	Speech Detector	Video+Speech Score Vectors	Video Score Vectors
Outdoors	.59	.58	.58	.58
Indoors	.12	.07	.23	.18
Face	.17	.15	.21	.18
People	.18	.18	.24	.25
Cityscape	.31	.34	.32	.30
Landscape	.19	.14	.18	.18
MAP	.26	.24	.29	.28

Table 1: In-Basis Concept AP & Overall MAP

Concept	Best Video Detector	Video+Speech Score Vectors
Outdoors	.59	0.63
Indoors	.12	0.25
Face	.17	0.14
People	.18	0.27
Cityscape	.31	0.31
Landscape	.19	0.17
MAP	.26	0.29

Table 2: Out-of-Basis Concept AP & Overall MAP

with a multimodal basis set improves performance by 12% over our pre-deployed video-only detectors; for comparison, a score vector approach using a basis of only video score vectors improves MAP by 8%. The model-score space approach does not hurt (in fact, helps) for the in-basis case.

Task “OUT-OF-BASIS” and Results: The second experiment 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 2 shows MAP a little above our best video-only TREC 2002 detectors. The model-score-space approach has constructed useful models for new out-of-basis concepts.

5. CONCLUSIONS

This paper describes a general framework for incorporating multimodal cues into a system for automatically annotating arbitrary semantic concepts in broadcast video. Early results show the multimodal score-space concept detection framework improves MAP by 12% over our best video-only feature-space detectors for detecting user concepts of interest which occur in the pre-deployed basis set; competitive MAP is also obtained for concepts outside the pre-deployed basis set. Section 3 mentioned many issues remaining open, including basis set size and content and whether the approach will scale to large sets of user-defined concepts. An interesting side observation from Section 4 is that (at least for the Video TREC corpus) speech-based models perform surprisingly well for detecting traditionally “video-centric” concepts like “face”, “outdoors” and “cityscape”. Overall results represent a first step towards a user-trainable video annotation system using multimodal cues.

6. ADDITIONAL AUTHORS

C.-Y. Lin, M. Naphade, A. Natsev, C. Neti, J.R. Smith and B. Tseng, also at IBM TJ Watson Research Center. We thank M. Franz for speech retrieval tools.

7. REFERENCES

- [1] IBM Multimodal Annotation Tool. <http://www.alphaworks.ibm.com/tech/multimodalannotation>.
- [2] W. Adams and al. IBM Research TREC-2002 Video Retrieval System. In *Proc. TREC Workshop*, 2002.
- [3] G. Iyengar and al. Semantic Indexing of Multimedia using Audio, Text and Visual Cues. In *Proc. ICME*, 2002.
- [4] C.-Y. Lin and al. VideoAnnEx: IBM MPEG-7 Annotation Tool for Multimedia Indexing and Concept Learning. In *Proc ICME*, MD, USA, July 2003. <http://www.alphaworks.ibm.com/tech/videoannex>.
- [5] S. Robertson and al. Okapi at TREC-3. In *Proc. TREC Workshop*, 1995.
- [6] A. Senior. Face and Feature Finding for a Face Recognition System. In *Proc. Second International Conference on Audio- and Video-based Biometric Person Authentication*, March 1999.