MULTIMODAL VIDEO SEARCH TECHNIQUES: LATE FUSION OF SPEECH-BASED RETRIEVAL AND VISUAL CONTENT-BASED RETRIEVAL

A. Amir†, G. Iyengar‡, C-Y. Lin§, M. Naphade§, A. Natsev§, C. Neti‡, H. J. Nock‡, J. R. Smith§, B. Tseng§

‡IBM TJ Watson Research Center, Yorktown Heights, NY, USA
§IBM TJ Watson Research Center, Hawthorne, NY, USA
†IBM Almaden Research Center, San Jose, CA, USA

ABSTRACT

There has been extensive research into systems for content-based or text-based (e.g. closed captioning, speech transcript) search, some of which has been applied to video. However, the 2001 and 2002 NIST TRECVID benchmarks of broadcast video search systems showed that designing *multimodal video search systems* which integrate both speech and image (or image sequence) cues, and thereby improve performance beyond that achievable by systems using only speech or image cues, remains a challenging problem. This paper describes multimodal systems for ad-hoc search constructed by IBM for the TRECVID 2003 benchmark of search systems for broadcast video. These multimodal ad-hoc search systems all use a *late fusion* of independently developed speech-based and visual content-based retrieval systems and outperform our individual speech-based and content-based retrieval systems on both manual and interactive search tasks. For the manual task, our best system used a query-dependent linear weighting between speech-based and image-based retrieval systems. This system has Mean Average Precision (MAP) performance 20% above our best unimodal system for manual search. For the interactive task, where the user has full knowledge of the query topic and the performance of the individual search systems, our best system used an interlacing approach. The user determines the (subjectively) optimal weights A and B for the speech-based and image-based systems, where the multimodal result set is aggregated by combining the top A documents from system A followed by top B documents of system B and then repeating this process until the desired result set size is achieved. This multimodal interactive search has MAP 40% above our best unimodal interactive search system.

1. INTRODUCTION

Multimedia information retrieval (including video search) has traditionally been approached independently by the text and spoken document processing community (which uses only degraded text, e.g. closed captioning, spoken words, visual text) and by the video processing community (which uses only visual information). It seems reasonable to hypothesize that robust solutions for multimedia retrieval might be more easily obtained by *multimodal video search techniques* which utilize all information in the component modalities of multimedia data (including images, speech and non-speech audio) rather than individual modalities alone. While some queries can be answered respectably using speech transcripts alone (e.g. "find stories on topic X") and others answered acceptably using images alone (e.g. "find basketball games", "baseball games" or "nuclear mushroom clouds"), it seems plausible that performance could be improved by search techniques exploiting cues in both text and image(s). An extreme illustration is the subset of queries which can only be answered by systems making use of multiple modalities (e.g. "find shots in which Yasser Arafat is speaking in front of the Wailing Wall"). This hypothesis about the potential gains achievable through multimodal search techniques represents our belief that the individual modalities carry highly complementary information and should therefore be exploited in tandem to improve search performance[1]. Despite this, and despite the extensive research which has been expended on systems for independent content-based or text-based retrieval, the 2001 and 2002 NIST benchmark tests of systems for searching broadcast video showed that the problem of integrating information from multiple modalities within a single multimodal video retrieval system remained a challenge: for example, for manual search, unimodal (typically speech-based) systems were amongst the top results (see e.g. [2], [3]).

This paper describes multimodal systems constructed by IBM for the TRECVID 2003 benchmark of search systems for broadcast video. These systems all use a late combination or *late fusion* of independently developed speech-based and visual content-based retrieval systems, as will be described. In contrast to the trend seen by multiple benchmark groups in 2001 and 2002, in which multimodal systems often performed less well than unimodal (e.g. speech-only) systems, these multimodal systems outperform our individual speech-based and content-based retrieval systems on both *manual* and *interactive* search tasks[2]. Paper organization is as follows. Sections 2 and 3 discuss the (independently developed and tuned) unimodal content-based and speech-based retrieval systems. Sections 4 and 5 discuss techniques for late integration of these unimodal systems into multimodal systems for manual and interactive search. Section 6 presents experimental results. The paper ends with conclusions and future work.

[1] Evidence of the usefulness of complementary information sources in developing robust solutions can be drawn from areas such as audio-visual speech recognition, in which the complementarity of multiple information sources has been exploited to obtain more robust solutions [1].

[2] In *manual* search, as specified by NIST guidelines, the user interprets the statement of information need and formulates a query. The user does not see the search corpus and gets exactly one attempt to launch the query on the search system. In *interactive* search, the user can interact with the system based on intermediate results. The user can refine the query, select and provide positive and negative feedback to the system and such. In both cases, guidelines state query formulation (and interaction, if applicable) must take less than 15 minutes.

2. CONTENT-BASED RETRIEVAL SYSTEMS

2.1. Manual Content- and Model-based System ("MC+MBR")

Our manual content- and model-based retrieval (CBR) is described in [4] and is only outlined here. The MCBR system supports search using a variety of low-level features (e.g. color, texture, edge) extracted from keyframes of shots at both global and regional levels. In addition, the system also supports search using a set of higher-level semantic concepts [5, 6, 7] that result from automatic annotation of keyframes using multimodal semantic concept detectors (such as faces, outdoors etc): a capability not commonly found in CBR systems. The semantic concept detectors can be used explicitly for semantic retrieval based on a weighted combination of specific concepts or they can be used implicitly in similarity-based retrieval using semantic model vectors, or vectors of confidences with respect to the set of concept detectors [8]. The combination of these capabilities allows a rich query formulation involving weighting on presence of semantic concepts as well as low-level features.

2.2. Interactive Content- and Model-based System ("IC+MBR")

Our interactive content-based retrieval (CBR) is also described in [4]. In interactive mode, the user can further refine the query, choose additional semantic concept models or low-level features and investigate different aggregation techniques for combining multiple search results, in order to refine the search results.

2.3. Manual Fully Automatic Content-based System ("MECBR")

Automatic query formulation is not a formal requirement for the TRECVID search task, but we found it highly desirable given the NIST 15 minute constraint on manual and interactive query formulation. For this reason, we developed algorithms for the selection of the best visual query examples, features to be used in image search, granularities of image search, and methods for combination of scores from multiple example image searches: each individually a challenging problem since the solution may not exploit any prior knowledge of the queries or the search set. We refer to our fully automatic CBR approach as Multi-Example Content-Based Retrieval (MECBR), since we automatically query content by specifying multiple visual query examples using only a single query iteration. MECBR attempts to mitigate some of the semantic limitations of traditional CBR techniques by allowing multiple query examples and thus a more accurate modeling of the user's information need. It attempts to minimize the burden on the user, as compared to relevance feedback methods, by eliminating the need for user feedback and limiting all interaction—if any—into a single query specification step. It also differs from relevance-feedback (RF) methods in that MECBR usually involves the execution and combination of multiple simple queries rather than the continuous refinement of a single query, as in most RF methods. The design of MECBR positions it as a lightweight alternative for modeling of low-level and mid-level semantic topics, including semantically/visually diverse topics as well as rare topics with few training examples (e.g., see [9]).

Our original MECBR formulation does not require prior training or use feedback but it does require the user to specify one or more query examples and a fusion method to be used in combining these per-example results[3]. This requirement must be removed if we are to use it as our fully automatic image retrieval system. It is a challenging problem to automatically select the best query examples and fusion methods fully automatically. Our solution is as follows. We use all provided example images (and keyframes from any video clips) but, to reduce sensitivity to noise and outliers, we categorize these examples into visually/semantically coherent *categories* or clusters (similar to [9]). We then perform multiple image retrievals, using weighted Boolean AND logic for fusion within categories and OR logic for fusion across categories. We treat each category as equally important in retrieval, irrespective of its size; within a category, though, the importance of an example is defined to be inversely proportional to its distance to the category centroid.

[3] Note that in some practical applications, an initial feedback step might be required to locate images to form an initial query. This is not required in the 2003 TRECVID scenario.

3. SPEECH-BASED RETRIEVAL SYSTEMS

3.1. Manual Fusion SDR System ("FSDR")

The fusion speech retrieval (SDR, "Spoken Document Retrieval") system is based upon the LIMSI-supplied ASR transcripts [10] and phonetic transcripts produced in-house. It extends the system in [4]. Closed captioning and video OCR are not used. The system ranks documents using a weighted linear combination of five separate speech retrieval systems: three OKAPI-based[11], systems (two based on the raw documents and indexing documents of 100 words in length, which differ slightly in the definition of documents and in the scheme used for mapping from document scores to shot scores after retrieval; one indexing a story-segmented version of the transcripts, such that documents do not overlap story boundaries), one soft-boolean system and one hybrid phonetic retrieval system which indexes phonetic transcripts but assigns scores on a per-word basis as in the soft-boolean scheme. The weight assigned to each system's scores is estimated to maximize retrieval performance (Mean Average Precision) on in-house queries for the development data set. For the benchmark search, OKAPI system queries were formulated by investigating retrieval for the TRECVID 2003 topics on the development data set; a similar procedure was used to produce lengthier queries for the soft-boolean system[4].

3.2. Interactive SDR System ("ISDR")

For practical rather than theoretical reasons, the ISDR system used only the soft-boolean search system component of the FSDR system. The purpose of our ISDR system was twofold: query refinement and shots elevation. No relevance feedback was used. The user enters an initial query, and then begins an iterative process of browsing a thumbnails result table, optionally listening to some of the retrieved shots, and then refining or modifying the query. Once the query is sufficiently refined, the user may mark individual shots in the obtained result list as "relevant" or "irrelevant". Upon saving the marked list to a file, all shots marked as relevant are elevated to the top of the list, the irrelevant ones are removed from the list, and the rest, unmarked shots, are left intact. This reranked list forms the ISDR result.

3.3. Manual Fully Automatic SDR System ("ASDR")

One fully automatic speech-based run based on a single OKAPI system was also submitted, for comparison against the fully automatic MECBR system. Queries for this run were formed by stripping prefixes such as "Find shots with/of" from NIST queries.

4. MULTIMODAL SYSTEMS FOR MANUAL SEARCH

4.1. Manual Linearly-combined System ("MLinear")

The first multi-modal run formed a query-dependent weighted linear combination of per-shot scores from two independent unimodal

[4] Following NIST guidelines, query formulation time was limited to less than 15 minutes.

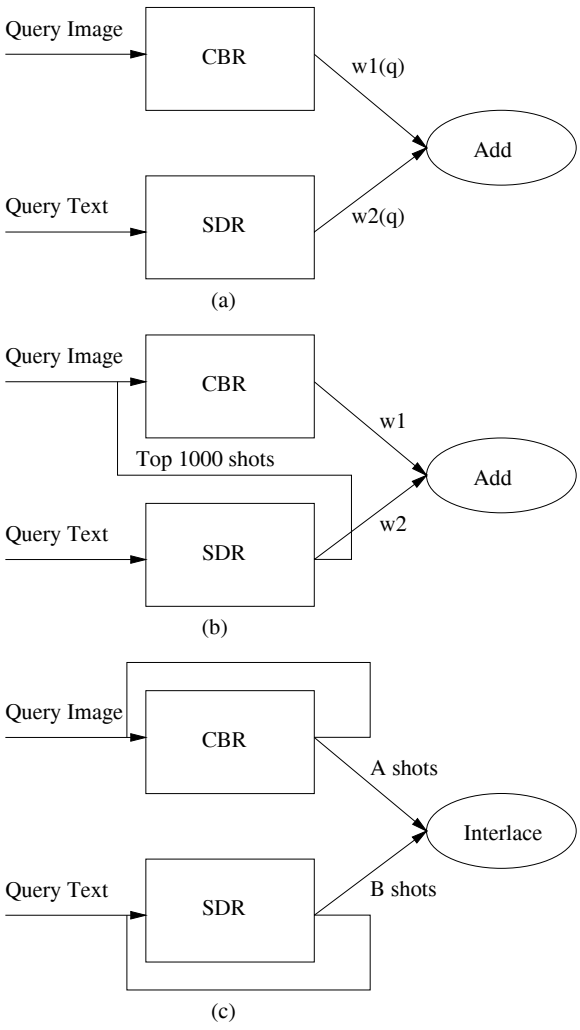

Fig. 1. Illustration of the 3 different fusion schemes investigated in this paper. (a) MLinear, (b) RSDR and (c) IFusion

runs (speech-based and content-based): specifically, FSDR and MECBR. This was in part a consequence of NIST test specifications: a maximum of 15 minutes is to be used in query formulation for the manual search task. Since we had already used about 14.5 minutes to explore speech-based query formulation on the development data, we chose to combine these results with those from MECBR: since the latter is fully automatic, the time-limit is not exceeded. This strategy was justified given the strong performance of speech systems in preceding evaluations. The per-modality weights were (for this system) set by the user, who uses their experience to predict which modality they expect to give better results. This modality receives a weight of 70% (30% for the other modality) and the target items are re-ranked using a weighted combination of the two individual scores. Note that whilst weighting on each modality is query dependent, it does not use search set knowledge or the actual performance of the systems (unlike an interactive run). Further analysis of the predictions made by users (their optimality, or lack of) will appear in a future work. Note also these were "expert" users; it remains to be seen whether novice users would make comparably successful predictions.

4.2. Manual Image-Reranked SDR System ("RSDR")

The second multi-modal run formed a query-independent weighted linear combination of per-shot scores from two independent unimodal runs (speech-based and content-based): specifically, FSDR and content-based retrieval. The speech-based retrieval was run first and only the top 1000 hits (i.e. shots) retained. A constrained global or regional content-based search was then performed, working only within this set of 1000 shots, using a user-selected "best" query image, region(s) and low-level features. The final score assigned to each of these 1000 shots was a weighted linear combination of the FSDR and the content-based retrieval scores. Shots outside the top 1000 received zero scores. The query-independent per-modality weights were (for this system) determined using the retrieval performance for the unimodal systems on in-house queries for the development data. (Note the contrast with `MLinear`, which uses query-dependent weights.). To illustrate, for the "Yasser Arafat" query, we refine the top 1000 FSDR results with a region-based retrieval using the colour and texture of Yasser Arafat's distinctive headscarf.

5. MULTIMODAL SYSTEMS FOR INTERACTIVE SEARCH

We chose to allow the user to spend the NIST-required maximum of 15 minutes interactively exploring using only one modality (i.e. CBR or SDR); we then combined these interactive results with the results of a fully automatic (zero user effort) manual search in the complementary modality, using a late fusion aggregation strategy. In other words, because we had generated fully automatic runs in both the speech-based and content-based modalities, we could use the entire allotted time exploring the search data using the more promising modality and to then improve results by combining them with an automatic run from the other modality at a zero cost with respect to the total user time spent for a given query. With the final weight determination strategy, the user has full knowledge of both the query topic and the performance of the two independent systems on that topic. Based on that knowledge the user determines the mixing weights A and B for the two modalities, where A+B=10. The multi-modal run is computed by interlacing the two result sets (sorted by relevance) so that the aggregated run takes the top A items from the first set followed by the top B new items from the second set, and so on; specifically, the interactive run for the higher-weighted modality is interlaced with the automatic run for the lower-weighted modality. We term this system **IFusion**.

6. EXPERIMENTAL RESULTS

We present results on the official TRECVID 2003 dataset, which comprises 170 hours of broadcast news content[5] from sources such as CNN Headline News, ABC World News Tonight and CSPAN. Approximately 63 hours of this content comprises development data, which we chose to use for system tuning. The queries used are the NIST-supplied 2003 test set of 25 multimedia query topics. These comprise textual statements of information need (e.g. "Find shots of Yasser Arafat") plus image and video snippet exemplars. The evaluation metric used is non-interpolated Mean Average Precision (MAP), a convenient single number for use in comparing systems. For a particular query, the average precision (AP) of a system is the sum of precisions at each correctly retrieved document in the result set divided by the total number of relevant doc-

[5]This dataset shares some of the content used for HUB4 and TDT evaluations.

uments in the entire corpus for that query. AP can be visualized as the area under the normalized precision-recall curve, with 1 being the area of the ideal curve. MAP is the mean of average precisions for all the queries.

Table 1 shows results for the systems discussed earlier, indicating type of system (manual, interactive or automatic), modality (or modalities) used (speech, visual or multimodal) and MAP performance for the TRECVID 2003 queries[6].

System	Type	Modality	MAP
MECBR	Automatic	Visual	0.04
MC+MBR	Manual	Multimodal	0.05
ASDR	Automatic	Speech	0.09
FSDR	Manual	Speech	0.12
RSDR	Manual	Multimodal	0.12
MLinear	**Manual**	**Multimodal**	**0.15**
ISDR†	Interactive	Speech	0.14
IC+MBR†	Interactive	Multimodal	0.13
IFusion†	**Interactive**	**Multimodal**	**0.20**

Table 1. Summary of retrieval evaluations of the various multimedia search systems using TRECVID 2003 corpus and queries

6.1. Manual Search Results

The table shows that whilst the best manual unimodal retrieval system is speech-based (FSDR), the best manual multi-modal system improves the MAP score over the respective constituent component single-modality systems by 20% for manual search. We note that the multimodal RSDR system does not improve compared to the unimodal FSDR system. Upon further examination, we note that while there was significant improvement for some queries (e.g. the "basketball", the "baseball" and the "Yasser Arafat" queries) using the RSDR system, the choice of a global weight between modalities as opposed to a query-dependent weight seems to have hurt its overall performance. Further analysis of this particular system will appear in a subsequent publication.

6.2. Interactive Search Results

Interactive outperforms manual within system design: both IC+MBR (MAP = 0.13) and ISDR (MAP = 0.14) results are higher than the corresponding manual runs (MC+MBR, MAP = 0.05; manual soft-boolean SDR, the comparable SDR system which is also a subcomponent of the FSDR system, has MAP = 0.09). In the IFusion system, where the user determines mixing weights after observing the performance of the component systems, we note that the performance is 40% better in terms of MAP score.

6.3. Other Observations

Our best multimodal manual retrieval system (MLinear, MAP = 0.15) performs comparably to the ISDR and IC+MBR systems. MC+MBR bought us little over plain MECBR on this year's data. This may be due in part to the difficulty of predicting the semantic concepts associated with a given query, when performed without the benefits of interactivity (which significantly increases performance).

7. CONCLUSIONS

This paper has described techniques for multimodal video search developed by IBM for the TRECVID 2003 benchmark. The techniques are all variants of late fusion between content-based and speech-based retrieval. The results support our initial hypothesis that use of complementary multimodal information can improve search performance on both manual and interactive tasks. Further evidence that search can be improved through multimodal techniques can be found in the results of other TRECVID 2003 participants. CMU Informedia reports up to 22% improvements from multimodal systems relative to a (degraded-)text baseline (manual task); the Lowlands Team reports a 1-2% gain from multimodality (manual task) (see [12]). Taken together though, these results suffice to show - on a large and standardized benchmark - that multimodal search techniques show promise for pushing the state-of-the-art in video search.

[6] †NIST results were later modified to average across only 24 queries for the interactive task, but we use 25 queries for comparison with the manual results.

8. ACKNOWLEDGEMENTS

We thank Paul Over, Alan Smeaton, Wessel Kraaij and NIST for organizing the TRECVID benchmark, Winston Hsu at Columbia University for story segmentation, Larry Sansone for in-house queries and IBM HLT volunteers for ground-truthing queries.

9. REFERENCES

[1] G. Potamianos, C. Neti, G. Gravier, and A. Garg, "Automatic recognition of audio-visual speech: Recent progress and challenges," *Proc. IEEE*, vol. 91, no. 9, September 2003.

[2] AF Smeaton and P Over, "The TREC-2002 Video Track Report," in *Proc TREC-2002*, MD, USA, 2002.

[3] AF Smeaton et al, "The TREC-2001 Video Track: Information Retrieval on Digital Video Information," in *Proc. ECDL 2002*, Italy, 2002, LNCS2458.

[4] W. H. Adams, A. Amir, C. Dorai, S. Ghosal, G. Iyengar, A. Jaimes, C. Lang, C-Y Lin, A. Natsev, M. Naphade, C. Neti, H. Nock, H. H. Permuter, R. Singh, J. R. Smith, S. Srinivasan, B. L. Tseng, A. T. Varadaraju, and D. Zhang, "IBM Research TREC-2002 video retrieval system," in *Proceedings of Eleventh Text REtrieval Conference(TREC-11)*, E. M. Voorhees and D. K Harman, Eds., Gaithersburg, MD, 2003, NIST, vol. SP500-251.

[5] W. H. Adams, G. Iyengar, C-Y Lin, M. R. Naphade, C. Neti, H. J. Nock, and J. R. Smith, "Semantic indexing of multimedia content using visual, audio and text cues," *EURASIP Journal on Applied Signal Processing*, , no. 2, pp. 170–185, February 2003.

[6] M. Naphade, C-Y. Lin, A. Natsev, B. Tseng, and J. R. Smith, "A framework for moderate vocabulary semantic visual concept detection," in *Intl. Conf. Multimedia and Expo*, Baltimore, MD, July 2003.

[7] C-Y. Lin, B. Tseng, M. Naphade, A. Natsev, and J. R. Smith, "Videoal: A novel end-to-end mpeg-7 video automatic labeling system," in *Intl. Conf. Image Processing*, Barcelona, Spain, Sept. 2003.

[8] J. R. Smith, M. Naphade, and A. Natsev, "Multimedia semantic indexing using model vectors," in *Intl. Conf. Multimedia and Expo*, Baltimore, MD, Jluy 2003.

[9] Apostol Natsev and John R. Smith, "Active selection for multi-example querying by content," in *Intl. Conf. Multimedia and Expo*. July 2003, IEEE.

[10] J. L. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system," *Speech Communication*, vol. 37, no. 1-2, pp. 89–108, 2002.

[11] S.E. Robertson, S. Walker, S. Jones, M.M. Hancock-Beaulieu, and M. Gatford, "Okapi at trec-3," in *Proceedings of the Third Text REtrieval Conference (TREC-3)*. 1995, pp. 109–126, NIST Special Publication 500-226.

[12] *TRECVID Workshop Notebook Papers*, MD, USA, 2003.