

Investigation of the Effectiveness of Cross-Media Indexing

No Author Given

No Institute Given

Abstract. Cross-media analysis and indexing leverages the individual potential of each indexing information provided by different modalities, such as speech, text and image, to improve the effectiveness of information retrieval and filtering in later stages. The process does not only constitute generating a merged representation of the digital content, such as MPEG-7, but also enriching it in order to help remedy the imprecision and noise introduced during the low-level analysis phases. It has been hypothesized that a system that combines different media descriptions of the same multi-modal audio-visual segment in a semantic space will perform better at retrieval and filtering time. In order to validate this hypothesis, we have developed a cross-media indexing system which utilises the *Multiple Evidence* approach by establishing links among the modality specific textual descriptions in order to depict topical similarity.

1 Introduction

A major challenge lies in developing representations suitable for crossing media and languages in the processes of *retrieval*, *filtering*, *categorization* and *summation*. The process of cross-media indexing consists of building relationships among concepts extracted from different modalities such as image, speech and text while reducing their weaknesses. It has been hypothesized that a system that combines different media descriptions of the same audio-visual segment will perform better at retrieval and filtering time. A robust indexing model could reconcile and recognize the relationships among concepts identified in these individual modalities and build not only a unified representation but also enrich the descriptions. Therefore, a cross-media analysis system should aim at minimising uncertainty, imprecision and inconsistency across the indexing performed by the single modalities. The main objective of the EU-IST Reveal-This (RT) project is to design, develop, test a complete umbrella infrastructure that will integrate a whole range of information access technologies across media and languages. A critical objective of the project is to benefit from multi-modal analysis and indexing techniques that can extend the possibilities of content push and pull technologies and increase the quality of content provided. In order to address the challenges mentioned previously and leverage the potential of different modalities, we have developed a prototype cross-media indexing system as part of the project. The prototype implements a standard way of identifying,

accessing, manipulating, storing and retrieving semantic links between modalities. The prototype system constitutes the last layer of the chain of processing in the Cross-Media Analysis and Indexing subsystem of the RT system. It utilises the *Multiple Evidence* approach by establishing links among the modal descriptions in order to depict topical similarity in the semantic textual space. The prototype’s input is a merged representation of the high-level features that are extracted by the speech processing component (speaker turns, transcriptions, persons, topics), the text-processing component (named entities, terms, facts), the image processing component (faces, key-frames, visual features), the image and the text categorizers. The media files were automatically segmented to form stories. After this stage, the descriptions are merged. This representation is further processed by the our prototype to produce a unified view of the content and enrichment.

In this study, we want to know whether cross-media indexing using *Multi Evidence* approach is effective or not. We admit that, neither cross-media indexing is an easy task nor such a question can be answered quickly. Therefore, we break down the question into several testable hypotheses and investigate them one by one.

We state our *generic* hypothesis as whether *Multi Evidence* approach performs better over a baseline system given the processed test collection, queries and graded relevance judgments. We also state *specific* hypotheses as follows:

1. No modality is dominant over other modalities, which means every modality has a significant contribution for effective retrieval.
2. Imprecision in the speech modality (recognition errors) can be remedied by other modalities for effective retrieval.
3. Document expansion may remedy the imprecision of speech modality.

We first gave an overview of the cross-media indexing task, the project and the prototype system as well as our research questions. In the rest of the paper, we focus on the detailed evaluation of our approach and discussion of the results. The next section will provide a review of the related work and highlight some research issues. Section 3 gives inside about the theory that is used for cross-media indexing. In section 4, we describe the evaluation strategy that is followed for our hypotheses and the prototype. Later in this section, we provide results from our experiments, discuss the issues we have encountered and provide further experimentation in various directions. The last section summarises key points of the paper and explores future directions.

2 Related Work

Indexing and searching for multimedia units constitute a great challenge due their inherent diversity and composite nature. Unlike text, other modalities such as audio (i.e speech) and image are difficult to work with. The diversity of modalities require dissimilar examination techniques specific to their characteristics.

Cross-media¹ analysis and retrieval research is an highly inter-disciplinary field and has recently received attention of various other research communities. It is the product of recent advances in speech, image, text analysis and retrieval research and many others. In its widely studied form, the subject matter of the field is composite content such as audio-visual segments or even multimedia rich web sites. Together with the TRECVID initiative and thus the availability of a test collection, numerous works have been done reporting varying results on the assumptions of the research area.

The research done so far can be grouped into four broad categories. The first one can be defined as the study of feature extraction techniques; how they are extracted, how low-level features are mapped to their high-level correspondences, whether retrieval in each modalities (other than text) can be effectively improved. For a long time, individual modalities have been examined extensively in their own domain² on several areas. The research in speech retrieval, for example, paid a lot of attention in late 90's for improving the effectiveness of spoken document and video retrieval systems. It was reported that [2, 3], even under high levels of word recognition error rates, it is possible to get reasonable retrieval performance using classical IR techniques. A review of the speech retrieval field is presented in [2]. The author argues that the accuracy of spoken document retrieval systems is not adversely affected by speech recognition errors by providing evidences from the prior work. He connects this argument to the results of repetition of important words in the output text and the results of additional related words providing a greater context. Based up on his review, he suggests that there is a minimum number of words necessary for the redundancy and context effects to overcome problems due to ASR errors. He also emphasises that previous work on ASR and IR together helped developing robust weighting schemes, in techniques that should be able to cope with misspellings, and with the idea that it makes sense to expand a document rather than a query. In [4], the authors focus on expanding documents to alleviate the effect of transcription mistakes on speech retrieval. A discussion of the effects of out of vocabulary items in spoken document retrieval is given in [5]. Woodland shows that, the experiments on TREC-8 (1998) audio collection using various retrieval setups suggests that it is possible to demonstrate moderate retrieval performance when advanced IR techniques (a combination of query and Document expansion method) are used to compensate for recognition errors caused by out of vocabulary words. The experiments include query expansion and document expansion against a baseline retrieval system that uses OKAPI variant. The second research issue is how and in which level these features are combined or *fused* [6, 7]. The recent work in the image retrieval research reports that text (in the form of annotations or speech transcripts) and image (histogram, texture, colour frequencies etc.) can be combined successfully [8, 9]. In [8], the authors suggest an approach for predicting words that are associated to whole images and corresponding to particular image regions. The process is regarded as a problem of translation,

¹ Notion of *Multi-modal* is often used interchangeably

² An up-to-date survey of the research fields can be found in [1]

such as translation of image regions to words. The work proposes an extension to Hofmann’s [10] hierarchical clustering/aspect model.

The third issue is to what extent a single modality or any combination of the modalities can improve the effectiveness of retrieval [11, 12]. In [13], Duygulu extends the previous work and studies videos rather than single images. This is a difficult problem since, text is not associated with a single frame. The authors report reasonable success over TRECVID-2001 collection. In another study by [14], semantic labeling is formulated as a machine learning problem. Concept representations are modeled using Gaussian mixture models, Hidden Markov Models, and Support Vector Machines and were evaluated against TRECVID-2001 collection. The study reports that fusion scheme achieves more than 10% relative improvement over the best single modal concept detectors.

The fourth research issue is users. User’s interaction with multimedia objects, information need formulations, and how appropriate representations can be decomposed and interpreted by the system is the focus of this area. Admittedly, certain modalities (such as text and speech) are more suited for information need expression than others. There has been less work done in conducting user studies with regard to assessing the effectiveness of the cross-media hypothesis. In [15], the authors report on TRECVID-2003 interactive search task by comparing three systems’ performances (text only, feature only, combined). According to the findings, the system which combined both text and other modal features did not perform well as expected. Furthermore, the experiments revealed that concept-based video retrieval worked best for *specific* topics, and hybrid system performed better on *general* topics. The study also gives emphasis on the importance of task descriptions and granularity of the annotation of data collection which is assumed to have considerable effects on the experiments.

3 Indexing Model

The prototype utilises Dempster-Shafer’s Theory of Evidence [16] approach for establishing links among the modal descriptions in order to depict *topical similarity* in the textual space. The theory has been extensively studied in image retrieval [17–19] and structured document retrieval [20], but has never been applied in such a context.

Dempster’s work lays the foundations of a non-Bayesian theory of probability which was then extended by Shafer who introduced the evidence combination rule. The theory combines two or more bodies of evidence defined within the same *frame of discernment* T into one body of evidence. We used the work of Lalmas [20] in a slightly different way. In our application of the model, we consider different modalities as sources of evidences. A document d containing a term t ’s existence in a modality m is counted as an *evidence* to support the topical similarity hypothesis. Therefore, each modality is treated as a probability density function also called as *Base Probability Assignment (BPA)*.

$$m_d(\emptyset) = 0 \quad \text{and} \quad 1 = \sum_{t \in d} m(\{t\}) . \quad (1)$$

where

$$m_d(\{t\}) = \begin{cases} tf(d, t) \cdot \log_N(\frac{N}{n(t)}) & \text{if } t \in d, \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

$$m_d(T) = 1 - \sum_{t \in d} m_d(\{t\}) . \quad (3)$$

In practice, the flexibility of the framework allows us to replace the formula in equation 2 by another probability density function. The $tf(d, t)$ component of the expression stands for the term t 's probability which is calculated as the number of occurrences of term t divided by the number of all terms in the document d . The \log component is a variant of inverse document frequency where N represents the number of documents in the collection and $n(t)$ represents the number of documents that contain term t . The preceding formula in equation 2 gracefully adheres to the theory by $\sum_{t \in d} m_d(\{t\}) \leq 1$ and therefore $m(T)$ is equal to the unassigned BPA in equation 3 when $m_d(\{t\})$ is smaller than 1. We combine evidences from modal descriptions by applying the following combination rules:

$$m(\{t\}) = m_1 \otimes m_2(\{t\}) = \frac{1}{K} (m_1(\{t\}) \cdot m_2(\{t\}) + m_1(\{t\}) \cdot m_2(T) + m_2(\{t\}) \cdot m_1(T)) . \quad (4)$$

$$m(T) = m_1 \otimes m_2(T) = \frac{1}{K} (m_1(T) \cdot m_2(T)) . \quad (5)$$

where

$$K = \left(\sum_t m_1(\{t\}) \cdot m_2(\{t\}) \right) + m_1(T) \cdot m_2(T) . \quad (6)$$

such that $m_1(\{t\}) > 0$ and $m_2(\{t\}) > 0$ conditions are satisfied. The retrieval function is simply the sum of combined masses of the query terms those appear in the documents.

$$RSV_{d \in D} = \sum_{t \in Q \text{ and } t \in d} m(\{t\}) . \quad (7)$$

4 Evaluation

Evaluating the accuracy and robustness of individual processing modals should be done in their own domain and with specially crafted test beds. However, when it comes to evaluating the performance of the merged and enriched representation of these processors, one needs to find a solution. In an attempt to validate our arguments and thus our prototype system, we considered two different testing strategies; *known item search* and *task oriented user test*, both involving the construction of a test collection using real users. For building such a collection one would need to have a set of “documents” (these are video/radio

segments which we call stories), a set of queries and a set of documents that have been found to be relevant to the queries (having exhaustively gone through the collection and found that these were all relevant documents contained in the collection).

4.1 Building of the Test Collection

For the purposes of the project we collected TV and radio broadcast parallel content in English and Greek languages, approximately 30+30 hours that covers news(40%), politics(21%) and travel(35%) domains. The collection contains multi-modal documents of different genre that guarantees a significant variety in modality-specific characteristics.

We reserved a subset of four-hours English version of the content for generating cross-media representations, queries and relevance judgments to be used in testing specifically our cross-media indexing prototype (see Table 1). Despite being relatively small, our intention was to have a data collection completely annotated with graded expert relevance judgments. Admittedly, manual annotation of every single segment is a labour and time intensive task especially when using non-binary relevance judgments. This is one of trade-offs of building a test collection one would confront. Due to the same reasons we were not able to manually create transcriptions and annotate feature in different modalities. Contribution of the manual transcripts and feature annotation are valuable to make a data set complete. These would provide us an established data set for comparison with other work for improving and evaluating individual and corporate performances. The lack of these led us to consider using other well known collections that can be acquired to test some of our other hypotheses (see section 4.3).

Table 1. Description of the collection used for testing Cross-Media Indexing prototype

The subset of the test collection			
<i>Domain</i>	<i>Medium</i>	<i>Duration</i>	<i>Stories</i>
Politics (EU Plenary sessions and Press Conferences)	Radio	-	-
	Video	01:09:58	31
News	Radio	-	-
	Video	00:44:26	33
Travel	Radio	00:28:12	24
	Video	03:01:44	188
Video sub-total		03:46:10	252
Radio sub-total		00:28:12	24
Grand total		04:14:23	276

Nevertheless, considering the characteristics of the modalities and the content of our collection, we created a set of 87 queries. Three experts were involved in the

annotation task. They were asked to go through the collection exhaustively for each query and manually annotate categories and provide relevance judgements. The experts were also asked to justify their judgements when considering stories as “partially relevant” to indicate how strict/lenient they had been in their judgements. The segments that were not mentioned at all were assumed to be judged as irrelevant by the annotators.

For the analysis of the media files (subset collection), we have defined a process workflow so that everytime one of the processors is improved, a new version of the collection could be regenerated. This approach enabled us to resolve dependencies amongs the processing modules and more importantly test the effects of relative improvements of individual processors and their contribution to the overall enriched representation.

4.2 Retrieval Experiments

In order to assess the performance of our model, initially we used Vector Space Model and *tf-idf* weighting scheme as the baseline system. Both systems were run using the query set over the processed content on every single modality and over multiple modalities. Figure 1 depicts the precision vs. recall values from our experiments utilising multiple modalities.

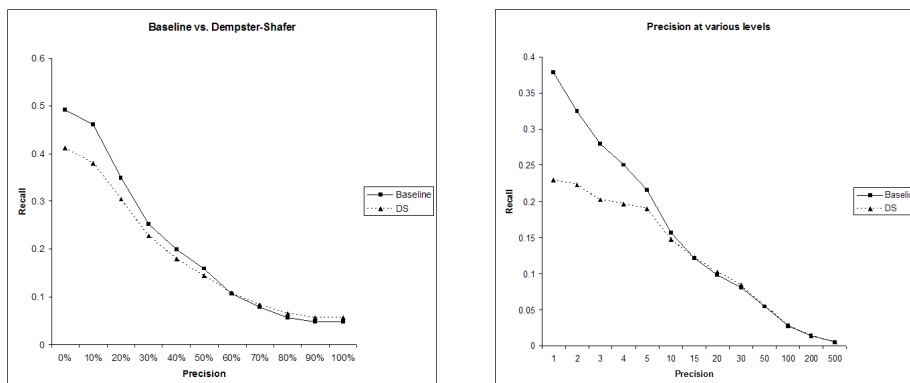


Fig. 1. Precision vs. Recall values for Baseline and Dempster-Shafer systems

The results suggest that Dempster-Shafer approach is performing significantly worse than our baseline system. Before drawing any quick conclusions and rejecting our general hypothesis, we wanted to make sure whether root cause of the problem was our proposed model. Despite being better than our model, the precision and recall values for the baseline system does not seem very high either. This may also be the result of factors such as the size of the collection or the imprecision of individual modalities. Further analysis of our empirical result also showed that there is no positive performance contribution of modalities

that are other than speech by which we reject our first and second hypothesis. Although not significantly, in some cases it was observed that modalities other than speech altogether had decreased the performance in both baseline and DS system. In other words, running both systems using speech only content yielded better performance. Given our setup and experiments, we conclude that speech modality seems to be dominant over other modalities. This might be due to three reasons in addition to the above mentioned ones: 1) Test queries might be biased towards speech 2) Size of the vocabulary is larger in this modality 3) Reducing every single modality to textual space may not be appropriately capture relations.

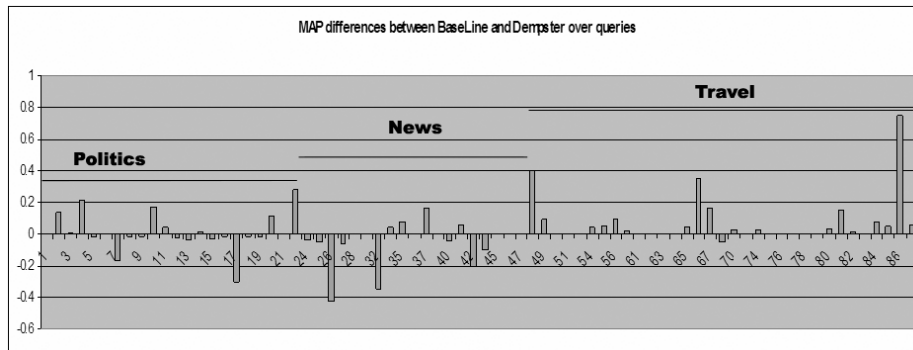


Fig. 2. Difference in MAP values between Baseline and Dempster-Shafer over all queries

In order to understand the effects of queries (at least partially), we analysed the Mean Average Precision (MAP) values of both systems to see for which queries the performance of the prototype decreases or increases. Fig. 2 illustrates the difference in MAP values for each query. The domains where the queries belong to are also indicated. The positive values show that baseline system's performance is better than DS for the specific queries. This is clearly the case for the queries belonging to the travel domain. Although not conspicuous at first sight, DS performs slightly better for politics domain. From our prior knowledge we know that, politics domain has totally different acoustics, the speech processor module is well trained in this domain and not very well in others. Therefore, it is not surprising to see speech processor doing better job in politics than news and travel. However, DS seems to be competitive in news domain as well. This diversity may well be one of the reasons for not catching up with baseline.

4.3 Discussion

Given that speech is dominant over all other modalities and there is no significant contribution from them, we returned to our third specific hypothesis. Admittedly, ASR (D_{asr}) is less accurate than manual transcripts (D_{mt}). As mentioned before, document expansion techniques can be employed to reasonably repair the ASR where the expanded representation ($D_{asr_{exp}}$) is closer to the original document (D_{org}) representation. Therefore, ideally a retrieval model would return the same document if we have an ASR representation which is closer to original document. We formalise this assumption as $D_{org} = D_{mt} \gg D_{asr_{exp}} \gg D_{asr}$. However, we also acknowledge that document expansion models should be used carefully as they may insert terms to the document which might move up non-relevant documents higher in the ranking.

4.4 Document Expansion from Web

Singhal et.al [4] makes several assumptions for using document expansion for Spoken Document Retrieval and advises that using a parallel corpus is beneficial for obtaining good results. Given the purposes and dimensions of our project, acquiring a parallel corpus which has the same period of time and diversity of domains is an unrealistic assumption. However, in order apply document expansion with some corpus that might show some resemblance in topicality, we decided to use WWW. Initially, Google search engine results are utilised to expand each story in the collection. Unlike running ASR documents as queries themselves [4], we selected top 10 representative and discriminative terms to be used as queries from the speech modality of the stories. The selection of terms was done by a slightly different algorithm using Kullback-Leibler Divergence measure applied in [21].

KL is typically used for measuring the difference between two probability distributions [22]. When applied to the problem of measuring the distance between two term distributions as in Language Modeling [23], KL estimates the relative entropy between the probability of a term t occurring in the actual collection Θ_c (i.e. $p(t|\Theta_c)$), and the probability of the term t occurring in the estimated Topic Language Model Θ_d (i.e. $p(t|\Theta_d)$). KL is defined as,

$$KL(\Theta_d||\Theta_c) = \sum_{t \in V} p(t|\Theta_d) \log \frac{p(t|\Theta_d)}{p(t|\Theta_c)} \quad (8)$$

where,

$$p(t|\Theta_c) = \frac{n(t, \Theta_c)}{\sum_{t \in \Theta_c} n(t, \Theta_c)} \quad (9)$$

and,

$$p(t|\Theta_d) = \frac{n(t, d) + \alpha}{\sum_{t \in d} n(t, d) + \alpha} \quad (10)$$

The expression $n(t, d)$ is the number of times term t occurs in a document d . The sparsity problem within the Θ_d is handled by Laplace smoothing. A non-zero constant α is introduced to alleviate the zero probability [23]. The smaller the KL divergence the closer the document is to the actual collection. A zero KL score indicates two identical distributions.

In our case, we are interested in each term t 's contribution to the KL score for a document d , instead of determining the difference between two term distributions. The greater the contribution to the document model the higher the KL score will be. Therefore, for each term t 's the contribution is calculated as in the following:

$$KL_d(t) = p(t|\Theta_d) \log \frac{p(t|\Theta_d)}{p(t|\Theta_c)} \quad (11)$$

The top 20 ranked terms in a document model are then ranked further according to each terms representative (generality) and discriminative (specificity) properties. The first $p(t|\Theta_d)$ part in the equations 11 determines the representativeness and the later \log component determines the discriminativeness. This approach makes it possible to measure the effectiveness of the two categories of terms in addition to our main objectives. Top ten qualifying terms are then used as queries to be submitted to the search engines. The contents of the first ten links were gathered and each story's speech content was expanded using the formula suggested by [24] and applied in [4] without additional weights. We continued our experiments using the base index (BI), base index extended by representative queries (IRQ) and base index extended by discriminative queries (IRD). Fig. 3 depicts the precision and recall values for baseline and DS over BI, IRQ and IDQ indexes. Although not significantly different than the baseline, it is observed that DS system shows some improvement over the IRQ index on the left hand side graph.

In addition to these experiments, we had also indexed ASR only and its expanded versions using Terrier system to run other 8 retrieval models [25]. We found out that there is no significant performance improvement for each retrieval model over these three indexes (F=1.694, Significance=0.184). The results also suggest that none of the retrieval models perform significantly better than the other ones (F=0.80, Significance=0.999).

5 Conclusions and Future Work

In this paper, our approach to cross-media indexing and initial experiments were presented. Given the available parameter space it is quite early to say that Dempster-Shafer approach is not working. The hypothesis of cross-media indexing and the models in use are still an open research area. We are on the way to validate our hypotheses. There are many issues, variables and problems yet to be tackled for a reliable system performance. There might be other cross-media models that can combine different modalities in their own feature space or equalize them in one modality such as text.

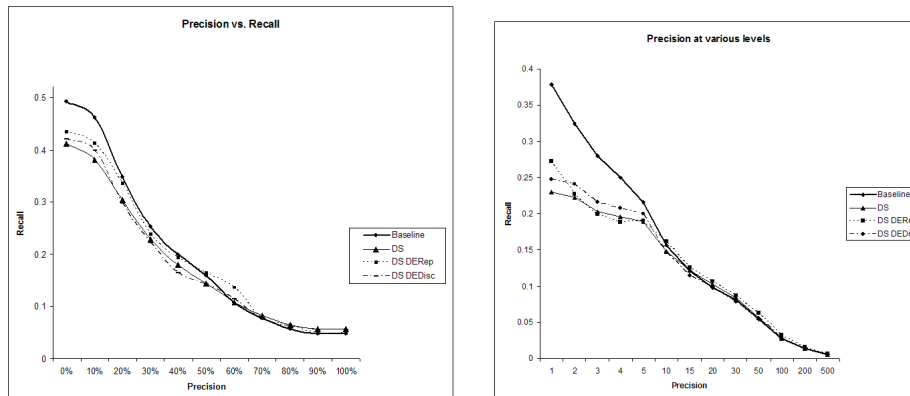


Fig. 3. The graph on the left illustrates the precision and recall values for Baseline, DS and DS expanded with Representative and Descriptive queries. The graph on the right hand side illustrates the precision values in various levels.

We will be working on various ways to improve the processes such as “cross-media” query generation and document expansion models provided in this study. We are also extending the experimentation to the use of different collections with various other search engines. In order to further investigate the effects of document expansion models using web, we acquired TDT-2 collection. Performing the same experiments using another collection will enable us to draw conclusions regards to the robustness of using web for this purpose.

References

1. Snoek, C., Worring, M.: Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications* **25**(1) (2005) 5–35
2. Allan, J.: Perspectives on information retrieval and speech. In: *ACM SIGIR Workshop: Information Retrieval Techniques for Speech Applications*, Springer-Verlag (2001) 1–10
3. Crestani, F.: Spoken query processing for interactive information retrieval. *Data and Knowledge Engineering* **41**(1) (2002) 105–124
4. Singhal, A., Pereira, F.: Document expansion for speech retrieval. In: *Proceedings of ACM SIGIR, Berkeley, California, United States*, ACM Press (1999)
5. Woodland, P.C., Johnson, S.E., Jourlin, P., Jones, K.S.: Effects of out of vocabulary words in spoken document retrieval (poster session). In: *Proceedings of ACM SIGIR Conference, Athens, Greece*, ACM Press (2000)
6. Wu, Y., Chang, E.Y., Chang, K.C.C., Smith, J.R.: Optimal multi-modal fusion for multimedia data analysis. In: *MULTIMEDIA '04: Proceedings ACM Multimedia*, ACM Press (2004) 572–579
7. Amir, A., Iyengar, G., Lin, C.Y., Naphade, M., Natsev, A., Neti, C., Jock, H.J., Smith, J.R., Tseng, B.L.: Multimodal video search techniques: late fusion of speech-based retrieval and visual content-based retrieval. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing. Volume III.* (2004) 1048–51

8. Barnard, K., Duygulu, P., de Freitas, N., Forsyth, D., Blei, D., Jordan, M.I.: Matching words and pictures. *Journal of Machine Learning Research* **3** (2003) 1107–1135
9. de Vries, A.P., Westerveld, T., Ianeva, T.: Combining multiple representations on the TRECVID search task. In: *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2004)*. (2004)
10. Hofmann, T.: Learning and representing topic. a hierarchical mixture model for word occurrence in document databases. In: *Workshop on Learning from Text and the Web*, CMU (1998)
11. Souvannavong, F., Merialdo, B., Huet, B.: Improved video content indexing by multiple latent semantic analysis. In: *CIVR'04, International Conference on Image and Video Retrieval*, *Lecture Notes in Computer Science*. Volume 3115., Springer (2004)
12. Snoek, C., Worring, M., Hauptmann, A.: Learning rich semantics from news video archives by style analysis. *ACM Transactions on Multimedia Computing, Communications and Applications* (in press)
13. Duygulu, P., Ng, D., Papernick, N., Wactlar, H.: Linking visual and textual data on video. In: *Workshop on Multimedia Contents in Digital Libraries*, Crete, Greece (2003)
14. Adams, W.H., Iyengar, G., Lin, C.Y., Naphade, M.R., Neti, C., Nock, H.J., Smith, J.R.: Semantic indexing of multimedia content using visual, audio, and text cues. *Journal on Applied Signal Processing* **2003**(2) (2003) 170–185
15. Yang, M., Wildemuth, B., Marchionini, G.: The relative effectiveness of concept-based versus content-based video retrieval. In: *Proceedings of ACM Multimedia*, ACM Press (2004) 368–371
16. Shafer, G.: *A Mathematical Theory of Evidence*. Princeton University Press (1976)
17. Jose, J.M., Harper, D.J.: A retrieval mechanism for semi-structured photographic collections. In: *Proceedings of the DEXA*, Springer-Verlag (1997)
18. Aslandogan, Y.A., Yu, C.T.: Multiple evidence combination in image retrieval: Diogenes searches for people on the web. In: *Proceedings of ACM SIGIR*, Athens, Greece, ACM Press (2000)
19. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: *Proceedings of ACM SIGIR*, Toronto, Canada, ACM Press (2003)
20. Lalmas, M., Moutogianni, E.: A Dempster-Shafer indexing for the focussed retrieval of a hierarchically structured document space: Implementation and experiments on a web museum collection. In: *Proceedings of RIAO*, Paris, France (2000)
21. Baillie, M., Elswiler, D., Nicol, E., Ruthven, I., Sweeney, S., Yakici, M., Crestani, F., Landoni, M.: University of strathclyde: the i-labs first big day out at trec hard. In Voorhees, E.M., Buckland, L.M., eds.: *Proceedings of the Fourteenth Text REtrieval Conference (TREC-14)*, NIST Special Publication (2005)
22. Kullback, S.: *Information theory and statistics*. Wiley, New York (1959)
23. Xu, J., Croft, W.B.: Cluster-based language models for distributed retrieval. In: *SIGIR '99: Proceedings of ACM SIGIR*, New York, NY, USA, ACM Press (1999) 254–261
24. Rocchio, J.J.: Relevance feedback in information retrieval. In Salton, G., ed.: *The SMART retrieval system: experiments in automatic document processing*, Englewood Cliffs, NJ, USA, Prentice-Hall (1971) 313 – 323
25. Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C.: Terrier: A High Performance and Scalable Information Retrieval Platform. In: *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*. (2006)