Multimedia Search with Pseudo-Relevance Feedback

Rong Yan, Alexander Hauptmann and Rong Jin

School of Computer Science
Carnegie Mellon University
Pittburgh, PA, USA
{yanrong,alex,rong}@cs.cmu.edu

Abstract

We present an algorithm for video retrieval that fuses the decisions of multiple retrieval agents in both text and image modalities. While the normalization and combination of evidence is novel, this paper emphasizes the successful use of negative pseudo-relevance feedback to improve image retrieval performance. While the results are still far from perrfect, pseudo-relevance feedback shows great promise for multimedia retrieval in very noisy data.

The Informedia Digital Video Library Project

Video is a rich source of information, with aspects of content available both visually and acoustically. The Informedia Digital Video Library project focuses specifically on information extraction from video and audio content. Since 1995, over two terabytes of online data have been collected, with automatically generated metadata and indices for retrieving videos from this library. The architecture for the project was based on the premise that real-time constraints on library and associated metadata creation could be relaxed in order to realize increased automation and deeper parsing and indexing for identifying the library contents and breaking it into segments. Library creation was an offline activity, with library exploration by users occurring online and making use of the generated metadata and segmentation. The goal of the Informedia interface was to enable quick access to relevant information in a digital video library, leveraging from derived metadata and the partitioning of the video into small segments.

The Informedia research challenge is how much can the video and audio be analyzed automatically and then made to be useful to a user. Broadly speaking, the Informedia project wants to enable search and discovery in the video medium, similar to what is widely available for text. One prerequisite for achieving this goal is the automated information extraction and metadata creation from digitized video. Once the metadata has been extracted, the system enables full-content search and retrieval from spoken language and visual documents. The approach that was ultimately successful was the integration of speech, image and natural language processing techniques for library creation and exploration. While much of the Informedia project has focused on interactive tools and techniques for finding relevant video clips in a large digital video collection, this paper will discuss the automated processing and retrieval techniques implemented in Informedia.

Video Retrieval from Mixed Text and Image Queries

In this paper, we present an algorithm for the task of video retrieval. A query, consisting of both text and images or video is posed against a video collection, and relevant shots are to be retrieved. Our system accomplishes this by fusing the retrieval results of multiple retrieval agents. The overall system can be decomposed into several agents, including a text-oriented retrieval agent (responsible for the text in speech transcripts, and Video OCR), a video-information oriented agent (responsible for searching manually provided movie abstracts and titles), and a basic nearest neighbor image matching agent which can be combined with classification-based Pseudo-Relevance Feedback (PRF). The classifier based PRF approach is intended to augment the image retrieval result by feeding back initial retrieval information and then refining the retrieval result. More details about this will be provided in the next section. Because the various agents produce retrieval results ranked in their own output scoring metrics, the decisions of all these agents are converted into a representation of posterior probabilities in an attempt to create a uniform output space. The linear combination of posterior probability is then applied to generate the final retrieval ranking decisions.

Image retrieval with classification pseudo-relevance feedback

Example-based image retrieval task has been studied for many years. The task requires the image search engine to find the set of images from a given image collection that is similar to the given query image. Traditional methods for content-based image retrieval are based on a vector model. These methods represent an image as a set of features and the difference between two images is measured through a (usually Euclidean) distance between their feature vectors. While there have been no large-scale, standardized

evaluations of image retrieval systems, most image retrieval systems are based on features such as color, texture, and shape that are extracted from the image pixels.

In our system two kinds of low-level features are used for finding similar images: color features and texture features. The color features are the cumulative color histograms for each separate color channel, where the three channels are derived from the HSV color space. We use 16 bins for hue and 6 bins for both saturation and value. We generate a texture feature for each subblock of a 3*3 image tessellation. The texture features are obtained through the convolution of the subblock with various Gabor Filters. In our implementation, 6 angles are used and each filter output is quantized into 16 bins. We compute a histogram for each filter and generate their central and second-order moments as the texture features. We concatenate all the features into a longer feature vector for every image; i.e. one vector for all color features and one vector for all texture features. We use a simple nearest neighbor (NN) image matching algorithm on both color and texture to produce the initial similarity results. In a preprocessing step, each element of the feature vectors is scaled by the covariance of its dimension. We adopted the Euclidean distance as the similarity measure between two images.

Although nearest neighbor search is the most straightforward approach to finding the matching images, it suffers from two major drawbacks. First, irrelevant features in the vector are given equal weight to important features, and thus retrieval accuracy will hurt decrease dramatically. Feature selection is therefore a necessary step prior to computing the nearest neighbor images. In theory, relevance feedback, through re-weighting and query refinement, is a powerful tool to refine the feature weighting so as to provide more accurate results. However, it is impossible to obtain the user judgment information in most automatic retrieval tasks. A second negative aspect is the unjustified distance function. Since an appropriate distance measure is a function of both the characteristics of the dataset and of the queries, a simple Euclidean distance function is unlikely to work for all the queries and images. Another concern is the normalization of the different dimension of a feature vector. To mitigate all these issues, we propose a classification-based pseudo-relevance feedback approach to refine the initial retrieval result. Support Vector Machines (SVMs) are used as our basic classifier mechanism, since SVMs are known to yield good generalization performance compared to other classification algorithms.

The basic idea for this approach is to augment the retrieval results by incorporating the classification output value through Pseudo-Relevance Feedback (PRF). The input data for the classifier is based on the information provided by our initial retrieval results. Standard PRF methods, which originated in the text information retrieval community, utilize the top-ranked documents as positive examples to improve the accuracy. The idea is to re-weight the words in the document feature vector based on the words in the top ranked documents, which are assumed to be positive examples. However, due to the poor initial performance of current video retrieval system, even the very top-ranked results are not always the correct ones that meet the users' information need. Unlike in text retrieval methods, it is more appropriate to make use of the *lowest* ranked documents in the collection after the initial search, which are more likely to be the negative examples. Therefore, we construct a classifier where the positive data are the query image examples and the negative data are sampled from the least confident image examples in the initial retrieval results.

Since the number of positive examples in our retrieval task is always much smaller than the number of the negative examples, we cast the problem into the imbalanced dataset classification framework. To sample more negative examples but achieve an overall balanced distribution of negative and positive examples in the classifier training set, we apply an ensemble of SVMs to tackle the rare class problem. The overall procedure can be summarized as follows,

1. Generate the initial classification results by nearest neighbor retrieval for all the images in the collection.
2. Choose all the query images as positive data. Denote the number of query images as *m*.
3. Construct a negative sub-collection based on the initial retrieval results, which are defined by the lowest 10% of the retrieved data from the collection. We sample *k* groups of negative data from the negative sub-collection, where each group contains *m* query images. Combine each group of negative data and all the positive data as a training set.
4. Build a classifier from each training set to produce new relevant score for any images x $f_i(x)(1 \le i \le k)$, where i is the index of training set
5. Combine the results in form of logistic regression, which is

$$P(+ \mid x) = \frac{\exp(\beta_0 + \sum_{i=1}^{k} \beta_i f_i(x))}{1 + \exp(\beta_0 + \sum_{i=1}^{k} \beta_i f_i(x))}$$

In our system, we simply set β_0 as 0, $\beta_i (1 \le i \le k)$ as equal values.

Our approach presented here utilizes the collection distribution knowledge to refine the final result. Due to the good generalization ability of the SVM algorithm, the most relevant features are selected automatically. Also the approach yields a better distance function based on the probability estimation compared with the simple Euclidean distance.

Combination of multiple agents

As the first step to integrate different types of agents, all the relevance scores of the agents are converted into posterior probability. For each agent other than the classification-based PRF agent, the posterior probability is generated by a linear transformation of their rank and scaled to the range of [0, 1]. All these posterior probabilities are simply linear combinations as follows:

$$Score = a_I(b_c P_{color}(+|x) + b_t P_{texture}(+|x) + b_{PRF} P_{PRF}(+|x)) + a_T P_{text}(+|x) + a_m P_{movie}(+|x)$$

where a_I, a_T, a_m is the weight for image agent, text agent, movie information agent respectively, which are set to be 1, 1, 0.2. b_c, b_t, b_{PRF} are the weights for the three search agents for image retrieval: NN on color, NN on texture and classification PRF, which are either set to be 0 or 1 in our contrastive experiments reported below.

Experiment

The video data came from the video collection provided by the TREC Video Retrieval Track. The definitive information about this collection can be found at the NIST TREC Video Track web site: http://www-nlpir.nist.gov/projects/trecvid/. TheText REtrieval Conference evaluations are sponsored by the National Institute of Standards and Technology (NIST) with additional support from other U.S. government agencies. Their goal is to encourage research in information retrieval from large amounts of text by providing a large test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results
The first Video Retrieval Track evaluation was performed in 2001, it's purpose was the investigation of content-based retrieval from digital video. The retrieval evaluation centered around the shot as the unit of information retrieval rather than the scene or story/segment as the video document to be retrieved.

The 2003 Video Collection for the Video Trec retrieval task consisted of ~40 hours of MPEG-1 video in the search test collection. This translated to 1160 segments as processed by Carnegie Mellon University, or 14,524 shots where the boundaries were provided as the common shot reference of the Video TREC evaluation effort. The shots were composed of a total of 292,000 I-frames which we extracted directly from the MPEG-1 compressed video files.

The actual 25 queries can be categorized into four general types in the 2002 Video Retrieval task:

- **•Specific item or person**

  Eddie Rickenbacker, James Chandler, George Washington, The Golden Gate Bridge, The Price Tower in Bartlesville, Oklahoma, etc.

- **•Specific facts**

  The Arch in Washington Square Park in New York City, an image of a map of the continental United States, etc.

- **Instances of a category**

  Football players, overhead views of cities, one or more women standing in long dresses, etc.

- **•Instances of events/activities**

  People spending leisure time at the beach, one or more musicians with audible music, crowd walking in an urban environment, a locomotive approaching the viewer.

Speech Recognition

The audio processing component of our video retrieval system splits the audio track from the MPEG-1 encoded video file, and decodes the audio and downsamples it to 16kHz, 16bit samples. These samples are then passed to a speech recognizer. The speech recognition system we used for these experiments is a state-of-the-art large vocabulary, speaker independent speech recognizer. For the purposes of this evaluation, a 64000-word language model derived from a large corpus of broadcast news transcripts was used. Previous experiments had shown the word error rate on this type of mixed documentary-style data with frequent overlap of music and speech to be 35 – 40%.

Text Retrieval

All retrieval of textual material was done using the OKAPI formula. The exact formula for the Okapi method is shown in Equation (1)

$$Sim(Q,D) = \sum_{qw \in Q} \left\{ \frac{tf(qw,D)\log\left(\frac{N - df(qw) + 0.5}{df(qw) + 0.5}\right)}{0.5 + 1.5\frac{|D|}{avg\_dl} + tf(qw,D)} \right\} \quad (1)$$

where *tf*(*qw*,*D*) is the term frequency of word *qw* in document *D*, *df*(*qw*) is the document frequency for the word *qw* and *avg_dl* is the average document length for all the documents in the collection.

Results

We report our results in terms of mean average precision in this section, as shown in Table 1. Four different combination of the retrieval agents are compared in this table, including the combination of text agents (Text), movie agents (Movie), nearest neighbor on color (Color), nearest neighbor on texture (Texture) and classification-

based PRF (Classification). The results show a significant increase in retrieval quality using classification-base PRF technique. While the text information from the speech transcript accounts for the largest proportion of the mean average precision (0.0658), only a minimal gain was observed in the mean average precision when the 'movie title' and abstract were also searched (0.0724) in addition to the speech transcripts. The image retrieval component provided further improvements in the scores to a mean average precision of 0.1046. Finally, the PRF technique managed to boost the mean average precision to the final mean average precision score of 0.1124.

Approach	Precision	Recall	Mean Average Precision
Text only (ASR)	0.0348	0.1445	0.0658
Text + Movie information (Abstract and Title)	0.0348	0.1445	0.0724
Text + Movie + Image retrieval (Color + Texture)	0.0892	0.220	0.1046
Text + Movie + Color + Texture + PRF Classification	0.0924	0.216	0.1124

Table 1 Video Retrieval Results on the 25 queries of the 2003 TREC video track evaluation.

Conclusions

We present an algorithm for video retrieval by fusing the decisions of multiple retrieval agents in both text and image modalities. While the normalization and combination of evidence is novel, this paper emphasizes the successful use of negative pseudo-relevance feedback to improve image retrieval performance. While the results are still far from satisfactory, PRF shows great promise for multimedia retrieval in very noisy data.

Acknowledgements

This work was partially supported by National Science Foundation under Cooperative Agreement No. IRI-9817496, and by the Advanced Research and Development Activity (ARDA) under contract number MDA908-00-C-0037.

References

Hafner, J. Sawhney, H.S. Equitz, W. Flickner, M. and Niblack, W. "Efficient Color Histogram Indexing for Quadratic Form Distance," IEEE Trans. Pattern Analysis and Machine Intelligence, 17(7), pp. 729-736, July, 1995.

Robertson S.E., et al.. Okapi at TREC-4. In The Fourth Text Retrieval Conference (TREC-4). 1993.

Sato, T., Kanade, T., Hughes, E., and Smith, M. Video OCR for Digital News Archive. In *Proc. Workshop on Content-Based Access of Image and Video Databases.* (Los Alamitos, CA, Jan 1998), 52-60.

Singh, R., Seltzer, M.L., Raj, B., and Stern, R.M. "Speech in Noisy Environments: Robust Automatic Segmentation, Feature Extraction, and Hypothesis Combination," *IEEE Conference on Acoustics, Speech and Signal Processing*, Salt Lake City, UT, May, 2001.

A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta and R. Jain, "Content-Based Image Retrieval at the End of the Early Years," IEEE Trans. Pattern Analysis and Machine Intelligence, 22(12), pp. 1349-1380, December, 2000.

Swain M.J. and Ballard, B.H. "Color Indexing," Int'l J. Computer Vision, vol. 7, no. 1, pp. 11-32, 1991.

Tague-Sutcliffe, J.M., "The Pragmatics of Information Retrieval Experimentation, revised," Information Processing and Management, 28, 467-490, 1992.

TREC 2002 National Institute of Standards and Technology, Text REtrieval Conference web page, http://www.trec.nist.gov/, 2002

The TREC Video Retrieval Track Home Page, http://www-nlpir.nist.gov/projects/trecvid/

Wactlar, H.D., Christel, M.G., Gong, Y., and Hauptmann, A.G. "Lessons Learned from the Creation and Deployment of a Terabyte Digital Video Library", *IEEE Computer* **32**(2): 66-73.

Informedia Digital Video Library Project Web Site. Carnegie Mellon University, Pittsburgh, PA, USA. URL http://www.informedia.cs.cmu.edu

A. Del Bimbo " Visual Information Retrieval", Morgan Kaufmann Ed., San Francisco, USA, 1999

Mojsilovic, J. Kovacevic, J. Hu, R.J. Safranek, and S.K. Ganapathy, "Matching and Retrieval Based on the Vocabulary and Grammar of Color Patterns," IEEE Trans. Image Processing, 9(1), pp. 38-54, 2000

Gong, Y. *Intelligent Image Databases: Toward Advanced Image Retrieval*. Kluwer Academic Publishers: Hingham, MA.