Multimedia Search with Pseudo-Relevance Feedback

Rong Yan, Alexander Hauptmann and Rong Jin

School of Computer Science
Carnegie Mellon University
Pittsburgh, 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. Although we have not solved all problems in video information retrieval, the results are encouraging, indicating that pseudo-relevance feedback shows great promise for multimedia retrieval with very varied and errorful data.

1 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 a text description plus 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, which is responsible for finding the text in the speech transcripts [4] and Video OCR [3], a video-information oriented agent which is responsible for searching the '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 motivation of the classifier based PRF approach is to improve the image retrieval performance by feeding back relevance estimates based on the initial search results into a classifier and then refining the retrieval result using the classification output. This approach will be described in more detail in the next section. To address the issue of comparability between retrieval scores produced by different types of agents, the retrieval scores of these agents are converted into posterior probabilities in an attempt to create normalized output scores. The posterior probabilities are then linearly combined to generate the final ranking decisions.

2 Image retrieval with classification pseudo -relevance feedback

2.1 Traditional image retrieval framework

Content-based image retrieval (CBIR) [1, 5] has been studied for many years. The task requires the image search engine to find a set of images from a given image collection that is similar to the given query image. Traditional methods for CBIR are based on a vector space model [14]. These methods represent an image as a set of features and the difference between two images is measured through a similarity function 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 representing color [6], texture, and shape that are extracted from the image pixels [12].

'Nearest neighbor' search is the most straightforward approach to find matching images. It contains the implicit assumption that for each feature the class posterior probabilities (distributions) are approximately constant [16] for matching and non-matching images. However nearest neighbor search suffers from two major drawbacks. First, the nearest neighbor might assign equal weight to both the relevant features and irrelevant features. Thus, if a large number of the features of an image are irrelevant to the query, the retrieval accuracy will suffer dramatically, since many images similar with respect to the irrelevant features will be preferred. It is therefore reasonable to select a subset of features or re-weight the features before the nearest neighbor search. However, most feature selection techniques require either large amounts of labeled data or knowledge of the way the classes are distributed with respect to each feature. Applying feature selection to the retrieval problem becomes rather difficult since usually only a small number of (image) query examples are given. Relevance feedback [15] is one effective way to gather information about the class distribution through iterative interaction with users. Through either feature re-weighting or query refinement, relevance feedback has been shown to be a powerful tool for providing more accurate retrieval results [15]. However, it is not possible to obtain user judgments in automatic retrieval tasks. The second drawback to nearest neighbor search is the fixed similarity metric. Since an appropriate similarity measure can vary with different datasets and different queries, any fixed similarity metric is unlikely to work well over all possible queries and data collections. To address this, Hastie et al. [16] have proposed an adaptive nearest neighbor algorithm in which the similarity metric can be locally adapted to the features relevant for each query point and globally optimized using dimensionality reduction. In a similar spirit, we observe that that some learning algorithms can model the data distributions even with very few training data points, so therefore we postulate that they are also feasible candidates to remedy the drawbacks of the nearest neighbor search.

2.2 A classification-based pseudo-relevance feedback approach

As noted before, nearest neighbor search suffers from a lack of adaptability. Recent studies have shown that some well-established classification algorithms [21] can yield better generalization performance than nearest neighbor type algorithms. Following this direction, we propose a classification-based pseudo-relevance feedback approach outlined below in an attempt to apply SVM ensembles to refine the initial retrieval result in content-based video retrieval.

2.2.1 Retrieval as classification with additional training examples from PRF

Quite naturally, information retrieval can be treated as a binary classification problem, where the positive data are the relevant examples in the collection and the negative data are irrelevant ones. However, information retrieval and classification have inherent differences. Typically, a retrieval algorithm might only obtain a small amount of training 'data' from the images that form the query, and even more crucially, there is no negative training data at all. To apply a classification algorithm to the retrieval problem, we have to provide more training data for the classifiers. One possibility is to identify the potential positive and negative class labels of unlabeled image examples in the collection with aid of the hints from initial search results. This is what we call pseudo-relevance feedback (PRF) in our work.

Therefore, the basic idea for our approach is to augment the retrieval performance by incorporating classification algorithms via PRF, with the choice of training examples based on the initial retrieval results. Standard PRF methods, which originated from the filed of text information retrieval [17], view the top-ranked documents as positive examples. Most text-based PRF methods update the feature weights based on the word frequency in the top ranked documents. However, due to the limited accuracy of current video retrieval systems, even the very top-ranked results are not always the relevant, correct answers that meet the users' information need, so they cannot be used as reliable positive examples for relevance feedback.

However, we discovered that it is quite appropriate to make use of the *lowest* ranked documents in the collection, because these documents are very likely to be negative examples. Therefore, after the initial search, we can construct a classifier to produce a more reliable retrieval score, where the positive data are the query image examples and the negative data are sampled from the least relevant image examples. The classification confidence is then merged with the initial retrieval scores as the final score.

From the viewpoint of machine learning, the approach presented here can be thought of as positive example based learning or partially supervised learning [18, 19]. It has been proved [18] that accurate classifiers can be built with sufficient positive and unlabeled data without any negative data. These results provide a sound theoretical justification for our approach. However, the goal of these learning algorithms differs from our approach, because they mostly aim at assigning examples into one of the given categories, instead of producing a ranked list of the examples. Another line of research related to this work is learning with a small set of labeled data. Transductive learning is one of the most popular paradigms to handle small numbers of labeled

data. Transductive learning has recently been successfully applied in the area of image retrieval [20]. However, that work is also different from ours because we do not have any certain negative labels at all.

2.2.2 SVM Ensembles

In our experiments, support vector machines (SVMs) [21] serve as our base classifier, since SVMs are known to yield good generalization performance compared to other classification algorithms. The decision function is of the form

$$y = sign\left(\sum_{i=1}^{N} y_i a_i K(x, x_i) + b\right)$$

where x is the d-dimensional feature vector of a test example, $y \in \{-1,1\}$ is a class label, x_i is the vector for the i^{th} training example, N is the number of training examples, $K(x, x_i)$ is a kernel function, $a = \{a_1,...,a_N\}$ and b are the parameters of the model. These a_i's can be learned by solving following quadratic programming (QP) problem,

$$\min \quad Q(a) = -\sum_{i=1}^{N} a_i + \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j y_i y_j K(x_i, x_j)$$

subject to $\sum_{i=1}^{N} a_i y_i = 0$ and $0 \le a_i \le C \forall i$

However in our case, since the number of negative data examples is often much larger than the positive data set, it might be a bad idea for SVMs to train on all of the positive and negative data at one time. This asymmetric data distribution often leads to a trivial SVM classifier that produces negative output for all possible input data. Recent years have seen several attempts at addressing the rare class problem using different techniques such as up sampling, down sampling [22], boosting [23] and biased discriminant analysis [25].

Along this direction, Yan et al. [24] have proposed a hierarchical classification solution by using SVM ensembles to tackle the imbalanced data set problem. To generate training sets with balanced distributions without either removing any training data or significantly increasing the training time, they first decompose the negative data into several partitions, and combine all the positive examples with each partition of negative examples to be an individual subset. A set of constituent SVMs is trained independently on every subset of the training set. Finally, all constituent SVMs will be combined by various strategies, including majority voting, sum of the posterior probability, and meta-classifiers (stacking). In this work, we adopt a similar framework, with the modification that a logistic regression algorithm is used to combine the output of the constituent SVMs.

2.2.3 Algorithm Details

The overall procedure of our image retrieval algorithm can be summarized as follows:

1. Generate the initial classification results by nearest neighbor search for *color* features of all the images in the collection.
2. Generate the initial classification results by nearest neighbor search for *texture* features of all the images in the collection.
3. Utilize all the query images as positive data. Let m be the number of query images.
4. Construct a negative sub-collection based on the initial retrieval results. In our implementation, the 10% least relevant images from the collection are chosen as the negative sub-collection. We sample k groups of negative data from the negative sub-collection, where each group contains m query images. Each group of negative data is combined with the positive data as a training set.
5. Build a classifier from each training set to produce new relevance score $f_i(x)(1 \le i \le k)$ for any images *x*, where *i* is the index of training set. $f_i(x)$ is set to 1 for positive prediction and 0 for negative prediction.
6. Combine the outputs of all the classifiers in the form of a logistic regression, which is

$$P_{PRF}(y=1|x) = \frac{\exp(b_0 + \sum_{i=1}^{k} b_i f_i(x))}{1 + \exp(b_0 + \sum_{i=1}^{k} b_i f_i(x))}$$

In our system, we simply set b_0 as 0, $b_i (1 \le i \le k)$ as equal values.

Our approach, as presented here, utilizes the collection distribution knowledge to refine the retrieval result. Due to the good generalizability of the SVM algorithm, the more relevant features are automatically more likely to be highly weighted. Also the approach yields a better similarity metric than the fixed one based on finding the largest margin between the positive and negative data examples for the current query.

3 Combination of multiple agents

Ultimately, the scores from multimodal agents have to be fused together to produce a final rank list for each query. The first step to integrate different types of relevance scores is to convert all the relevance scores into posterior probabilities. The conversion of PRF approach is described in section 2.2.3. For nearest neighbor type approached, the rank of each video shot is scaled to the range of [0, 1] by linear transformation. This normalized rank can be viewed as the posterior probability, which is, $P(y=1|x) = 1 - R/R_{\max}$ where $R_{\max}$ is the maximum rank.

All these posterior probabilities are simply linear combined to be the final score,

$$Score_I = b_c P_{color}(y=1|x) + b_t P_{texture}(y=1|x) + b_{PRF} P_{PRF}(y=1|x)$$

$$Score = a_I Score_I + a_T P_{text}(y=1|x) + a_m P_{movie}(y=1|x)$$

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

4 Experimental Results

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/. The Text 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. Its 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 2002 Video Collection for the Video TREC retrieval task consisted of ~40 hours of MPEG-1 video in the search test collection. The content came from the Internet Archive and Open Video websites, consisting mostly of documentary television from the 50's and 60's. This translated to 1160 segments as processed by Carnegie Mellon University's Informedia system [10, 11] or 14,524 shots where the boundaries were provided as the common shot reference of the Video TREC evaluation effort. We extracted a total of 292,000 I-frames directly from the MPEG-1 compressed video files to represent the video's visual information.

The actual 25 queries of the 2002 Video Retrieval track had an average of 2.7 video examples each, as well as an average of 1.9 image examples. The queries can be categorized into four general types:

- **Specific item or person**
  Eddie Rickenbacker, James Chandler, Abraham Lincoln, 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.

A sample query for images of *George Washington*, was represented in XML as:

```
<!DOCTYPE videoTopic SYSTEM
"videoTopics.dtd">
<videoTopic num="077">
<textDescription text="Find pictures of
George Washington" />
<imageExample src= "http:
//www.cia.gov/csi/monograph/firstln/955pres2.gif"
desc="face" />
<videoExample src="01681.mpg"  start="09m25.938s"
stop="09m29.308s" desc="face" />
```

The above query example contains both a still image as well as a short video clip of about 3.5 seconds depicting an image of a portrait of George Washington. A sample image from the collection video is shown at right.

4.1 Image Features

Two kinds of low-level image features are used by our system: color features [6] and texture features. We generate both types of features for each subblock of a 3*3 image tessellation. The color feature is comprised of the central and second-order color moments for each separate color channel, where the three channels come from the HSV (Hue-Saturation-Value) color space [26]. We use 16 bins for hue and 6 bins for both saturation and value. The texture features are obtained from 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 again generate their central and second-order moments as the texture feature. In total, we obtained 18 features for each subblock and concatenate them into a longer vector of 144 features for every image. 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.

4.2 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 down samples 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% [11].

4.3 Text Retrieval

All retrieval of textual material was done using the OKAPI formula [2]. The exact formula for the Okapi method is shown in Equation (1)

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

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. No relevance feedback at the text level was used.

As noted before, video information agent uses externally provided video information as another source to improve retrieval performance. For each query, the score of a video shot is set to 1 if any keyword of the query can be found in the video titles/abstracts for the corresponding movie, otherwise the score is set to 0.

4.4 Results

We report our results in terms of mean average precision in this section, as shown in Table 1. Four different combinations of the retrieval agents are compared in this table, including the combination of text agent (Text), video information agent for externally supplied movie titles and abstracts (VI), nearest neighbor on color (Color), nearest neighbor on texture (Texture) and classification-based PRF (PRF). The results show an increase in retrieval quality using the classification-based 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 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. Further experiments are needed to investigate how the various parameter settings and combination strategies affect the performance of PRF approach.

Approach	Precision	Recall	Mean Average Precision
Text only (Speech Recognition)	0.0348	0.1445	0.0658
Text + Video Information (VI)	0.0348	0.1445	0.0724
Text + VI + Color + Texture	0.0892	0.220	0.1046
Text + VI + Color + Texture + PRF	0.0924	0.216	0.1124

Table 1 Video retrieval results on the 25 queries of the 2003 TREC video track evaluation.

5 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. One of the future directions of this approach is to study the effect of different classification algorithms and explore better combination strategies than a simple linear combination of the individual agents.

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

[1] 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.

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

[3] 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.

[4] 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.

[5] 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.

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

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

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

[9] TREC Video Retrieval Track, http://www-nlpir.nist.gov/projects/trecvid/

[10] 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.

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

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

[13] 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

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

[15] Y. Rui, T. S. Huang, and S. Mehrotra, "*Content-based image retrieval with relevance feed-back in Mars*," in Proc. IEEE Conf. Image Processing, 1997, pp. 815-818.

[16] T. Hastie and R. Tibshirani. *Discriminant adaptive nearest neighbor classification and regression.* In David S. Touretzky, Michael C. Mozer, and Michael E. Hasselmo, editors, Advances in Neural Information Processing Systems, volume 8, pages 409--415. The MIT Press, 1996

[17] Carbonell, J., Y. Yang, R. Frederking and R.D. Brown, "*Translingual Information Retrieval: A Comparative Evaluation*," Proceedings of IJCAI, 1997

[18] Liu, B., Lee, W.S., Yu, P.S. and Li, X., *Partially Supervised Classification of Text Documents*, Proc. 19th Intl. Conf. on Machine Learning, Sydney, Australia, July 2002, 387-394.

[19] F. Denis. *PAC learning from positive statistical queries*. In ALT 98, 9th International Conference on Algorithmic Learning Theory, volume 1501 of Lecture Notes in Artificial Intelligence, pages 112-126. Springer-Verlag, 1998

[20] Y. Wu, Q. Tian, and T. Huang. *Discriminant-em algorithm with application to image retrieval*. In Proceedings to the IEEE Conference on Computer Vision and Pattern Recognition, volume 1, pages 222--227, June 2000. 12

[21] V.N. Vapnik (1995). *The Nature of Statistical Learning Theory.* Springer

[22] Foster Provost, "Machine Learning from Imbalanced Data Sets 101/1", *AAAI Workshop on Learning from Imbalanced Data Sets*, AAAI Press, Menlo Park, CA, 64-68

[23] M.V. Joshi, R.C. Agarwal, V. Kumar, " Predicting Rare Classes: Can Boosting Make Any Weak Learner Strong? ", *the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Edmonton, Canada, July 2003

[24] R. Yan, Y. Liu, R. Jin, A. G Hauptmann, "On Predicting Rare Class with SVM Ensemble in Scene Classification", To Appear in International Conference on Acoustics, Speech, and Signal Processing 2003, Hong Kong, China, April, 2003

[25] X. S. Zhou, T. S. Huang, " Small Sample Learning during Multimedia Retrieval using BiasMap" , in Proc. IEEE Conf. Computer Vision and Pattern Recognition, Hawaii, Dec. 2001

[26] O. Chapelle, P. Haffner and V. Vapnik, "*SVMs for histogram-based image classification* ", IEEE Transaction on Neural Networks, 9, 1999