Interactive Content-Based Retrieval Using Pre-computed Object-Object Similarities

Liudmila Boldareva and Djoerd Hiemstra

University of Twente, Databases Group,
P.O.Box 217, 7500 AE Enschede, The Netherlands
{L.Boldareva,D.Hiemstra}@utwente.nl

Abstract. We propose using truncated object-object similarity matrix as an access structure for interactive video retrieval. The proposed approach offers a scalable solution to retrieval and allows combination of different feature spaces or sources of information. Experiments were performed on TREC Video collections of 2002 and 2003.

1 Introduction

With rapid development of digital media in the past decade, *Content-based information retrieval* (CBIR) has become an active research area. Originally meant for text documents, information retrieval quickly became dearly needed for other media such as still images and video. Though CBIR usually suggests the retrieval of non-textual information, the term does not exclude text documents[1]. The goal of a retrieval system is to satisfy the *information need* of the user. The information need is communicated to the system, e.g. by providing an example query.

A number of approaches to CBIR exist. The pioneering image retrieval systems used large experience existing in the text retrieval domain, successfully adopting the vector space model [7,14,10]. Probabilistic approaches from text retrieval (e.g. [9,12] gained less popularity among non-text CBIR researches with some notable exceptions [3,19,16]. One of the reasons lies in the difficulty of translating the lower-level features into probability values. Other recent research is inspired by machine learning methods. Self-organising maps [11] and support vector machines [4,15] are employed to solve the problems of CBIR. Many existing retrieval systems rely on active participation of the searcher in the retrieval process, which is known as *relevance feedback* [13].

Regardless of the approach used, a retrieval system should be able to 'understand' the users' information need and provide him/her with satisfactory answers. The problem is that high-level content of a document, in the way a human being understands it, is hard to translate into a machine-language concept with current techniques for automatic lower-level feature extraction. Rich feature spaces might be created in an attempt achieve a correspondence between lower-level features and human perception. This immediately creates a

[1] In this paper the term 'document' is used in a broad sense, implying any source of information such as text, images, videos, etc.

P. Enser et al. (Eds.): CIVR 2004, LNCS 3115, pp. 308–316, 2004.

disadvantage—a high-dimensional space that is not well suited for fast access via indexing. It raises the scalability problem: methods that perform well on small collections can not be used on a collection of usable size, due to the 'dimensionality curse' [6]. In this paper we propose a framework for content-based indexing and retrieval, that

- is able to use any available technique for feature extraction, and allows easy combination of different sources of information;
- focuses on relevance feedback as an important component of the information retrieval process;
- allows efficient interaction with the user, i.e. it offers a solution to the scalability problem.

We present a description of the proposed framework in Sec. 2. Interaction between the system and the user is studied in Sec. 3. Experiments performed on the collection of the TREC Video retrieval workshop (TRECVID) [1] are presented in Sec. 4. Conclusions and future work directions can be found in the last section.

2 Probabilistic Indexing and Retrieval

Consider a collection $\mathcal{I}$ of information objects i among which there is one that the user is looking for, the *search target* denoted T [2]. During the search process, the system presents the user with intermediary retrieval results. The user can indicate which examples are *relevant* to his/her information need, those are *positive examples*. If an object is *not relevant* to the query, the user may indicate so, thus providing the system with *negative examples*. Given the feedback information, the retrieval system produces a new set of candidate documents to be assessed by the user. There may be several loops of *relevance feedback* during one search session.

We want to make use of the notions 'relevant' and 'non-relevant' without having to refer to lower-level (image) features. We do so by relating objects in the collection to each other. A binary variable δ_i, that takes values 1 and 0, denotes the events of *positive* and *negative feedback* respectively. For two documents the following reflects their 'measure of closeness': $P(\delta_i = 1|T)$, the probability of an object i marked by the user as relevant given that T can be referred to as the target for the search. When unambiguous, we use a shorthand notation $P(\delta_i|T)$.

2.1 Interactive Retrieval in a Probabilistic Framework

For interactive retrieval we use a probabilistic approach. The idea is to predict the set of documents relevant to the user's information need, based on his/her request, accompanied by feedback, and the data representation (i.e. our measure of closeness $P(\delta_i|T)$). Using Bayes' rule the problem can be stated as estimating the probability of relevance $P(T)$ given user's feedback $\delta^1, \ldots, \delta^n$ and the collection indexing [12,3,16].

[2] The search target may be a single document, but it can as well be a number of documents covering a certain subject satisfying the user's information need.

We write it down in the following iterative form, using the assumption that the $\delta^1, \ldots, \delta^n$ are conditionally independent given the target T :

$$P^{\text{new}}(T) = P(T|\delta^1, \ldots, \delta^n) = \frac{P^{\text{old}}(T) \prod_{s=1}^{n} P(\delta^s|T)}{P(\delta^1, \ldots, \delta^n)} \quad . \tag{1}$$

We distinguish the following factors that influence an interactive search session:

1. The input provided by the user who is assumed to be reasonable in his/her query formulation and feedback.
2. The current document representation. Within one search session, the indexing of the collection is a static component of the model.
3. The prior information about the relevance of documents in the collection.

Below we describe our approach to indexing of a multimedia collection.

2.2 Indexing: The Structure of the Association Matrix

Documents in the collection and their conditional probabilities $P(\delta_i|T)$ can be visualised as a directed graph with objects $i \in \mathcal{I}$ as nodes and arcs with weights $P(\delta_i|T)$ connecting them. In this way each object is *described* by its *associations* with a number of other objects linked to it. We call such representation of the collection an *association matrix*, denoted $\mathbb{M}$.

Ideally we want the associations to refer to high-level semantics (e.g. coming from users' judgements) which might not be achieved using lower-level features. Starting at the point when we do not have knowledge about the human perception of similarity, the associations need to be based on something different. We propose to bootstrap the process by basing the associations on a similarity measure on lower-level features, such as colour, texture, or shapes present in an image (e.g. as used in [7,16]). Typically such similarity measures take values in $\mathbb{R}$ or $\mathbb{R}+$ and thus cannot be directly used as an initial estimate for $P(\delta_i|T)$.

In our model we take pair wise similarities based on, e.g., pictorial features, and we are looking for an appropriate transformation to obtain probabilities. Any increasing function with the domain $\mathbb{R}$ and the range $[0, 1]$ could suit. When deciding the probabilities in our model, we would like to achieve equal emphasis of the alike similarities and obtain probabilities, uniformly distributed in $[0, 1]$. The used transformation spreads the observations evenly on this interval according to their probability of occurrence and not the magnitude of the similarity measure. As a result it reduces the influence of outliers and preserves the scale of the similarities between documents and 'improves the discrimination capabilities of the similarity measures' [2]. Since a priori we cannot prefer some documents of the collection to others in the sense of the distribution of $P(\delta_i|T)$, the underlying similarities are assumed to be random values conform to the same probability distribution—the normal distribution.

We transform the computed similarities by subtracting the sample mean and dividing by the sample standard deviation and then applying the standard normal cumulative distribution function, to obtain estimates of the probabilities

which are denoted by $P(\delta_i|T)$. The value of $P(\delta_i = 0|T) = 1 - P(\delta_i = 1|T)$ obtained in this way can be interpreted as a *P-value*, the probability that a variable assumes a value greater than or equal to the observed one strictly by chance [18]. Thus by specifying some α such that $P \leq \alpha$ only *significant* pairwise similarities and their corresponding $P(\delta_i|T)$ are taken into account, and the rest is replaced by an appropriate constant further denoted by $\bar{p}$. When updating $P(T)$ for each object in (1), $P(\delta_i|T)$ is substituted with $\bar{p}$ if it is below $1 - \alpha$. Here $1 - \alpha$ serves as a cut-off threshold for the right tail of the distribution. For the rest of the paper the corresponding threshold for the left tail is set to zero. A pair of documents i, T having their $P(\delta_i|T)$ significant, are called *neighbours*.

Keeping only neighbours for each element makes the association matrix sparse, which allows faster access to the data. In our experiments with an appropriate/optimal choice of the cut-off threshold, depending in particular upon the size of the collection, the association matrix can grow as slowly as linear without the loss of the search quality. Pre-computed probabilities allow easy combination of different modalities of otherwise hard to combine feature spaces, such as visual information from a shot and speech transcripts from spoken words [8].

3 Modelling Interaction for Retrieval

The user feedback. During the search session, the current probability of an element to be the user's search target $P(T)$ is updated according to (1). Every document can be either relevant to the user's information need or not, i.e. the events are disjoint: $P(\delta_i = 1|T) = 1 - P(\delta_i = 0|T)$. The objects that are not marked by the user as relevant, take part in the probability update as if they are explicitly rejected by the user.

To ensure that in the lack of positive examples, the excessive (implied) negative feedback does not bury the precious positive examples, the $\bar{p}$ is set in our experiments to a value in the interval $(0, \alpha)$, with the effect that in the ranked list of results the non-neighbours of negative examples do not precede the neighbours of known (if any) positive examples from the last iterations.

New display for the next iteration. Upon updating $P(T)$, a new set of objects should be presented for relevance judgement, to receive new evidence from the user. The display update is an important part of the search process, since efficiency and quality of retrieval depend on it. Each iteration should bring the user closer to his/her target object. 'Closer to the target' may have various interpretations, such as: the posterior probability $P(T)$ of the desired information object(s) tends to 1; or the target object approaches the top of the ranked list, etc. The goal of the search is not only to satisfy the users' need, but to do it in few iterations and/or in a limited amount of time. In this paper we report experiments performed with the following display update strategies.

Best-target strategy. Following *probability ranking principle* [12], $P(T)$ is treated as a score that the element receives during retrieval session. The next display set consists of (new) documents that have largest values of $P(T)$.

The Best-target strategy is plausible for the user unfamiliar with content-based retrieval (thus, the majority of potential users). The screen always contains objects that are the neighbours of good examples marked by the user. The user is able to observe the immediate result of his/her action. It is not clear however, whether this approach converges the search to the target quickly enough. Cox et al. [3] report that the Best-target search occasionally gets stuck in an isolated 'island' of non-relevant documents that are similar to each other only.

Non-deterministic strategies. The Randomised display set consists of objects picked from the collection at random. Uniform sampling may give relatively good representation of the collection, which supposedly allows to find the relevant documents quickly. Sampling could be especially useful at the beginning of a search session, when the system has little knowledge about the user information need. When, after a number of iterations, the mass $P(T)$ is concentrated on a small (relevant) subset of the collection, sampling the *whole* data becomes useless and may have negative effect on the search quality. To minimise this effect, Random-of-Best strategy makes the selection among those objects for which their probability to be the target increased since the last iteration, which are effectively neighbours of relevant examples, and/or not-neighbours of the non-relevant ones. Ideally, the number of elements of which $P(T)$ increases should shrink on to the group of documents that satisfy the user's information need.

4 Experiments and Evaluation

4.1 Interactive Experiment Setup

We use video data provided in the framework of `TRECVID`. The videos are segmented into shots, and from each shot a representative key frame is extracted. Conditional probabilities for the association matrix are estimated using a generative probabilistic retrieval model (see for detail [19]):

1. $\mathbb{M}^{\mathrm{v}}$ using on Kullback-Leibler divergence as similarity measure for Gaussian Mixture Models built on pictorial data;
2. $\mathbb{M}^{\mathrm{t}}$, using language model-based similarity on text from speech transcripts;
3. $\mathbb{M}^{\mathrm{vt}}$, a run-time combination of the two modalities, which adds up the relevance scores achieved in both matrices.

We conducted an empirical study on performance difference caused by the prior distribution of the probability of relevance. In order to provide a better than uniform prior probability of relevance, for a number of experiments the text from search topic descriptions serves as a query to match against the speech transcripts, using a language model [9]. In another version of the system the prior distribution is determined by the number of neighbours in the association matrix for each document, so that a document with many neighbours has higher chance to be displayed. This is useful when no prior information about the information need is available, for instance in un-annotated data, or when the query terms typed by the user, do not occur in the collection.

A retrieval session starts with browsing a display set of 12 key frames generated by the prior distribution of $P(T)$, which might be based on an initial text query. The user does not have to provide an example query image.

The documents in the ranked list are ordered by the decreasing probability of relevance. A standard TREC evaluation metric, mean average precision (MAP) is used as a measure of user's satisfaction (see [5, Appendix]). Where questionable, signed rank test is used to determine if a difference in performance between two methods is significant. If not stated otherwise, the significance level is $p \leq 0.05$

4.2 Automated Experiments

In the series of experiments, referred to as *automated* the user input has been replaced with relevance judgements available from TREC assessors who played the role of a 'generic user'. The experiments have been performed on a subset of the collection selected so that that half of the key frames was relevant to at least one of the 25 topics. The goal of such setup was to test the retrieval performance in our probabilistic framework, and to find optimum settings to be used in the experiments with real users.

Values in the association matrix. Values of MAP after each iteration using two types of the association matrix and their combination, with the best found values of $\bar{p}$, and two matrices with *all pairs* of probabilities, are plotted in Fig. 1. Combining visual and text modalities results in better performance than using either separately. In the runs where text from the topics description is used as the query (Fig. 1b), the difference in average precision is smaller, which is an expected result: the shots that are relevant because of the initial query text are put on top of the ranked list, and further search depends on this prior distribution by the nature of the Bayesian approach.

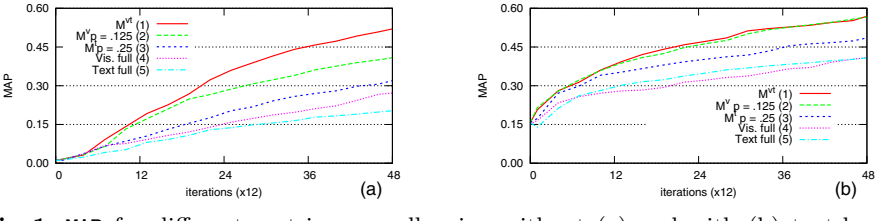

Fig. 1. MAP for different matrices vs. all pairs, without (a) and with (b) text-based prior distribution. In (b) the difference between curves (1) and (2) is not significant.

In the automated experiments the threshold $1 - \alpha$ is such that on average 3% of possible values need to be stored. Keeping only significant $P(\delta_i|T)$ in fact improves the search quality compared to the complete set of conditional probabilities both with visual and text-based matrices. This suggests that the probabilities replaced with the constant $\bar{p}$ are indeed far from true similarities.

The display update strategies. *The Best-target display update* with an ad-hoc tuned value of $\bar{p}$ offers great improvement over iterations, both when using the text priors and not. By making sure that the user does not see the same object twice, the danger of getting stuck in a local maximum is eliminated.

The two non-deterministic strategies perform not so well, especially when prior text information is used. The Non-deterministic methods perform on average 10 to 15 percent better if negative examples are *ignored* during the update of $P(T)$. As expected, the combination of Randomised and Best-target strategies (Random-of-Best) did better than the 'pure' Randomised. Still, uniform sampling of the more relevant subset of it, as done in Random-of-best, cannot beat the deterministic Best-target method. Sampling according to the estimated distribution of $P(T)$ might be a better option. In Fig. 2 the best-performing combination is plotted for each display update strategy.

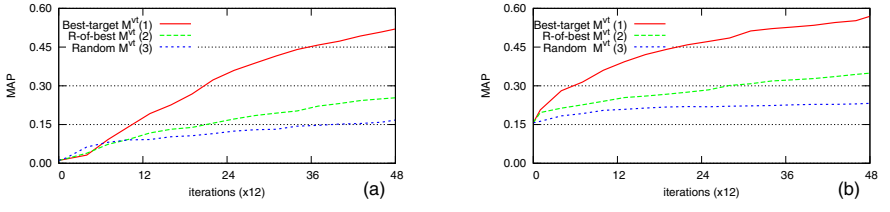

Fig. 2. MAP for different display update strategies without (a) and with (b) the prior text information.

The prior distribution based on text from the query description and words from speech transcripts provides overall better performance. Nevertheless, having little or no a priori information does not necessarily mean poor performance: Curve 1 in Fig. 1a for the method with no prior information available, reaches numbers comparable with the corresponding curve in Fig. 1b.

4.3 Live Experiments

In the *live* experiments, the search tasks have been performed by real users[3]. The data set contained about 32 000 key frames taken from 60 hours of news videos. We found high agreement between real users feedback and TREC relevance judgements (average among runs 75%), so our automated experiments can be viewed as a good approximation to real life (see [17] for an analysis of agreement between TREC assessors).

The set-up is similar to the automated experiments using the Best-target display update schema and text-based prior distribution $P(T)$. The user was allowed to see key frames (images), and not the corresponding videos. Only positive feedback from the user was taken into account. The resulting MAP at the

[3] 2 groups of 3 users to test 3 systems. All users are students of University of Twente aged between 19 and 26. Each search task took at most 15 minutes.

end of the live experiment evaluated by TREC is 0.245. For this run, 78% of the shots selected by the user were relevant according to TREC. At the same time, 48% of the relevant shots that have been displayed, were missed by our users. In the experiment that showed the user random screens (MAP 0.026), the number of missed shots was much lower (31%), as well as agreement with TREC (55%). The relevant documents are missed partially due to the fact that the user saw *still frames*, and not the *videos* themselves, but the difference in numbers between the runs suggests that relativeness of the users' judgements (the user selects best of what is available and two users do not always agree) plays a role, too.

4.4 Scalability of the Approach

The term 'scalability' denotes not only the possibility to run a retrieval system on a larger collection. The ability of a retrieval system to produce answers to the user's queries in a reasonable amount of iterations is at least as important.

We ran a number of automated experiments on a system consisting of 32 000 key frames from the TRECVID 03 data. After 48 iterations on the large collection, MAP of the best automated run is 0.44, compared to 0.58 achieved on the small collection. Note that half of the small collection were key frames relevant to one of the topics, whereas in the large collection only 6.5% of the key frames was relevant to one of the 25 topics. The execution time on the large collection, which is eleven times bigger than the small one, increased by factor 5 to 6.

5 Conclusions and Future Work

We found that feature normalisation and 'refinement' by way of replacing non-significant similarities with a constant which we propose, results in better search quality in both investigated feature spaces, text-based and visual-based. Using the association matrix as an index structure enables efficient combination of different modalities, such as visual information from key frames and transcripts of the speech occurring in video shots. Combining text and video (in the form of key frames) has positive effect on retrieval.

Organising the objects in a multimedia collection using the association matrix allows scalable implementation which is hard to achieve otherwise: computing similarities 'on the fly' is expensive in the sense of access time and/or computation effort, whereas keeping all pre-computed similarities is impractical from the storage point of view. Keeping only the significant similarities allows building an interactive content-based retrieval system that provides fast response time and good search quality on rather large image or video collections.

Text, in the form of speech transcripts of videos or annotations, is an important source of information about the multimedia content. When available, the text data should be used in combination with pictorial features, to improve the search results.

In the future we want to have the probabilities stored in the association matrix, to be updated by utilising the relevance judgements obtained from the

user's feedback. We are also going to investigate how to dynamically change the search strategy depending on user-system performance.

References

1. *TRECVID 2003 Workshop, Notebook Papers*, 2003.
2. S. Aksoy and R. M. Haralick. Feature normalization and likelihood-based similarity measures for image retrieval. *Pattern Recognition Letters*, 22(5):563–582, 2001.
3. I. J. Cox, M. L. Miller, T. P. Minka, T. V. Papathomas, and P. N. Yianilos. The bayesian image retrieval system, PicHunter: Theory, implementation, and psychophysical experiments. *IEEE Tran. On Image Processing*, 9(1):20–37, 2000.
4. H. Drucker, B. Shahrary, and D. C. Gibbon. Relevance feedback using support vector machines. In *Proc. 18th Int. Conf. on Machine Learning*, pages 122–129. Morgan Kaufmann, San Francisco, CA, 2001.
5. E.M.Voorhees, editor. *Proc. 10th Text Retrieval Conference, TREC-10*, 2002.
6. C. Faloutsos. *Searching Multimedia Databases By Content*. Kluwer Academic Publishers, Boston, USA, 1996.
7. M. Flickner, H. Sawhney, W. Niblack, and J. Ashley. Query by image and video content: the QBIC system. In *IEEE Computer*, volume 28, pages 310–315, 1995.
8. J. L. Gauvain, L. Lamel, and G. Adda. The LIMSI broadcast news transcription system. *Speech Communication*, 37(1-2):89–108, 2002.
9. D. Hiemstra. *Using language models for information retrieval*. PhD thesis, Centre for Telematics and Information Technology, University of Twente, 2001.
10. Y. Ishikawa, R. Subramanya, and C. Faloutsos. MindReader: Querying databases through multiple examples. In *Proc. 24th Int. Conf. Very Large Data Bases*, pages 218–227, 1998.
11. J. Laaksonen, M. Koskela, and E. Oja. PicSOM: Self-organizing maps for content-based image retrieval. In *Proc. of IJCNN'99*, Washington, D.C., USA, 1999.
12. S. E. Robertson. The probability ranking principle in IR. *Journal of Documentation*, 33(4):294–304, 1977.
13. J. J. Rocchio. Relevance feedback in information retrieval. In G. Salton, editor, *The Smart Retrieval System: Experiments in Automatic Document Processing*, pages 313–323. Prentice Hall, 1971.
14. Y. Rui and T. Huang. Optimizing learning in image retrieval. In *Proc. IEEE int. Conf. On Computer Vision and Pattern Recognition*, June 2000.
15. S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *Proc. 9th ACM Int. Conf. on Multimedia*, pages 107–118. ACM Press, 2001.
16. N. M. Vasconcelos. *Bayesian Models for Visual Information Retrieval*. PhD thesis, Massachusetts Institute of Technology, 2000.
17. E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing Management*, 36(5):697–716, 2000.
18. E. W. Weisstein. *CRC Concise Encyclopedia of Mathematics*. CRC Press, 2002.
19. T. Westerveld, A.P. de Vries, A.R. van Ballegooij, F.M.G. de Jong, and D. Hiemstra. A probabilistic multimedia retrieval model and its evaluation. *EURASIP Journal on Applied Signal Processing*, (2):186–198, February 2003.