Interactive retrieval of video using pre-computed shot-shot similarities

L. Boldareva and D. Hiemstra

Abstract: A probabilistic framework for content-based interactive video retrieval is described. The developed indexing of video fragments originates from the probability of the user's positive judgment about key-frames of video shots. Initial estimates of the probabilities are obtained from low-level feature representation. Only statistically significant estimates are picked out, the rest are replaced by an appropriate constant allowing efficient access at search time without loss of search quality and leading to improvement in most experiments. With time, these probability estimates are updated from the relevance judgment of users performing searches, resulting in further substantial increases in mean average precision.

1 Introduction

With the rapid development of digital media, content-based multimedia retrieval has become an active research area. Having started its history in text documents, information retrieval quickly became much needed for other media such as still images and video.

The pioneering image retrieval systems used experience from the text retrieval domain, successfully adopting the vector space model [1–3]. Probabilistic approaches suggested for retrieval [4, 5] gained less popularity with some notable exceptions [6–8]. One of the reasons for this lies in the difficulty of translating lower-level features that index the visual content into probability values. Often, content based retrieval systems rely on active participation of the searcher in the retrieval process, known as 'relevance feedback' [9]. Relevance feedback is a broad term sheltering various models of learning the user's information needs from the two-way communication where the searcher plays an active role. The early implementations of interactive visual retrieval systems are QBIC [1], MARS [3], MindReader [2], Viper [10], PicHunter [6]. More recently, machine learning methods have been applied successfully to visual information retrieval, e.g. self-organising maps [11] and support vector machines [12, 13].

One fundamental problem that needs to be solved in visual information retrieval, is the semantic gap—a mismatch between the human perception of visually rich documents and their representation in the storage. This has been an active research topic for decades [14–17]. We believe that simple tools based on low-level feature representations, and matching functions on them, are unlikely to bridge the semantic gap in the near future, without additional functionality like relevance feedback and long-term learning of image representations.

Another difficult problem in multimedia retrieval is the efficiency of search algorithms. High dimensionality of the search space hinders effective indexing of it, introducing the problem of 'dimensionality curse' [18]: with many dimensions, that feature vectors usually are, it is hard to implement a similarity matching algorithm that is substantially faster than a linear scan over the data [19]. This puts a possible interaction away. The system has to rely on dimensionality-reduction indexing techniques (e.g. [20]) or other smart approaches to access objects most similar to the given examples [21]. Still, the performance of such systems comes nowhere near to the performance of text retrieval systems. At the time of writing this paper, http://images.google.com provides access to almost 1200 million images using text search techniques, orders of magnitudes more than any content-based approach could possibly manage. The only viable solution may be processing as much of available information in advance as possible.

In this article we propose a framework for content-based indexing and retrieval, that

- can use any available technique for feature extraction and similarity matching, and allows easy combination of different sources of information;
- allows efficient interaction with the user, and is capable of learning from that interaction.

Therefore the proposed framework is spared both of the two problems stated above.

We give a statistical interpretation to the data-driven similarity between elements of the video key-frames collection that fits into a probabilistic framework for efficient interactive retrieval. This framework accommodates both short-term learning within one retrieval session and long-term learning from relevance feedback gathered in multiple retrieval sessions.

In the following Section a general Bayesian framework for interactive retrieval is introduced, and subsequently, two important components of it are discussed—the data representation and the user feedback. Section 5 discusses

IEE Proceedings online no. 20045201

doi: 10.1049/ip-vis:20045201

Paper first received 1st October 2004 and in revised form 12th May 2005

The authors are with the University of Twente, EEMCS faculty, P.O.Box 217, 7500 AE Enschede, The Netherlands

E-mail: boldarli@ewi.utwente.nl

long-term learning from previous searches based on user's relevance judgements. Experiments have been carried out to verify the benefits of the proposed methods and compare to other techniques. Data from the TREC Video retrieval workshop [22] serves as a testbed. The experimental set-up and results are reported in Section 7, after an overview of related work given in Section 6.

2 Interactive retrieval in Bayesian terms

Let $\mathcal{I}$ be a collection of information objects x, e.g. key-frames for video shots, among which there is what the user is looking for, the search target denoted here by T. During the search process, the system presents the user with intermediate retrieval results. The user can indicate which of the objects 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 objects to be assessed by the user. There may be several loops of relevance feedback during one search session.

The probabilistic framework is introduced as follows. Consider disjoint random events of the user feedback regarding the relevance of an object x. Let δ_x be the corresponding indicator function taking the value one if the user marks the candidate object x as relevant and zero otherwise. Note that the present framework can be generalised for multiple choices of feedback, e.g. by introducing a third event of explicit negative judgement.

We want to use the concepts 'relevant' and 'non-relevant' without having to refer to lower-level features. Instead, the objects in the collection are related to each other according to the most likely user opinion about their relationship. For two objects x and y, the following conditional probability reflects their 'measure of closeness': $P(\delta_x = 1|T = y)$, the probability of object x being marked by the user as relevant given that y is referred to as the target for the search. When unambiguous, the shorthand notation $P(\delta_x|T)$ denotes the probability of a certain user action concerning the object x.

2.1 *Estimating probability of relevance*

The goal is to predict, or identify, the set of objects relevant to the user's in formation need, based on his/her request accompanied by feedback and the existing data representation. In a Bayesian framework [5, 6] the problem is restated as estimation of the probability of relevance $P(T)$ given a user's relevance judgements $\{\delta_{x_1}, \ldots, \delta_{x_n}\}$ on the set of candidate objects $\{x_1, \ldots, x_n\}$ and the data indexing. We write it down in the following iterative form, with the assumption that the user actions $\{\delta_{x_1}, \ldots, \delta_{x_n}\}$ are conditionally independent given the target T.

$$P^{\mathrm{new}}(T) = P(T|\delta_{x_1}, \ldots, \delta_{x_n}) = \frac{P^{\mathrm{old}}(T)\prod_{s-1}^{n} P(\delta_{x_s}|T)}{P(\delta_{x_1}, \ldots, \delta_{x_n})} \quad (1)$$

The conditional independence assumption used here states that the user's judgement about the relevance of a certain item is not affected by the relevance of other displayed items.

$P^{\mathrm{new}}(T)$ becomes $P^{\mathrm{old}}(T)$ for the next iteration; $\{(\delta_{x_1}, \ldots, \delta_{x_n})\}$ are provided by the user. $P(\delta_{x_s}|T)$ represents conditional dependency between elements x and T. According to (1), in order to retrieve the meet likely relevant answers T, one needs to know all δ_{x_s} and $P(\delta_{x_s}|T)$. They are the subjects of the following two Sections.

3 Association-based data representation

If we were to use a graphic model, the collection of objects and the corresponding conditional probabilities $P(\delta_x|T)$ could be visualised as a directed graph with nodes $x \in \mathcal{I}$, and weighted arcs connecting them. Each object x is described by its associations with a number of other objects. The strength of the association is the weight of the arc, $P(\delta_x|T)$. The whole structure is called here 'association matrix' denoted by $\mathbf{M}$.

Preferably for each item there needs to be only a few associations, which refer to high-level semantics and agree with the observed users' acts of relevance feedback. We arrive at these associations as follows. Starting at the point when we do not have knowledge about human perception of similarities between objects, the initial associations are derived from a similarity measure on lower-level features, such as colours, textures, shapes. Typically such similarity measures take values in the range of (non-negative) real numbers and thus cannot be directly used as an initial estimate for $P(\delta_x|T)$.

As a first step, the pair-wise similarities are transformed by fitting a probability distribution on it. Since *a priori* we cannot prefer some objects from the collection to others in the sense of the distribution of estimates of $P(\delta_x|T)$, the underlying pair-wise similarities of the whole collection are assumed to conform to the normal distribution. This gives equal emphasis of the alike similarities and spreads the observations evenly on the interval [0, 1] according to their probability of occurrence and not to the magnitude of the similarity measure. As a result it reduces the influence of outliers and preserves the scale of the similarities between objects, i.e. 'improves the discrimination capabilities of the similarity measure' [23]. The result of this step is a square table with all possible pair-wisp estimates of conditional probabilities, also containing errors induced by the mismatch between data-driven similarity measures and human perception.

The value of $P(\delta_x = 0|T) \equiv 1 - P(\delta_x = 1|T)$ computed 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. That is, the P-value is the probability that the computed similarity between two objects purely by chance does not exceed the 'true' one, therefore x will not be found relevant to T by the user. As a second step, we specify some α, the upper bound for the P-value, so that only statistically significant pair-wise similarities and their corresponding $P(\delta_x|T)$ are taken into account. Probability estimates for not significant similarities are replaced by an appropriate constant further denoted by $\bar{p}$. That means, while updating $P(T)$ for each object in (1), the condition probabilities $P(\delta_x|T)$ is substituted by $\bar{p}$ if its estimate is below $1 - \alpha$. Here $1 - \alpha$ serves as a cut-off threshold for the right tail of the distribution of pair-wise similarities. Similarly, a threshold for the left tail of the distribution will differentiate between significant and non-significant estimates of dissimilarity. The corresponding threshold for the dissimilarity-based estimates for the rest of the article is set to zero, because judging non-relevance from low-level features is less practical. An object x that has a significant $P(\delta_x = 1|T = y)$ is called a neighbour of y. Our idea is that the estimates left out contain more noise than useful information and removing them will not harm retrieval quality, while improving efficiency.

The data representation in the form of association matrix as described here has the following advantages:

- different sources of information on the (estimated) relevance of objects are on the same scale and can be efficiently combined;

- the relevance information obtained from the searchers can be used to improve conditional probability estimates in the association matrix.

4 Input provided by the user

Another factor that plays a role in an interactive retrieval session is the input from the user. One question is how to interpret user actions. Another, related, question is how to select candidate objects to be presented to the user.

4.1 Interpretation of user feedback

During a search session, the current probability of an object to satisfy the user's information need $P(T)$ is updated according to (1). Every object can be marked relevant/non-relevant to the user's information need (or not marked at all), and these events are disjoint. If the user is not supposed to ignore the objects presented for relevance assessment, 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, and for their neighbours, $P(\delta_x = 0|T) \equiv 1 - P(\delta_x = 1|T)$ is used in (1) to update $P(T)$.

Deploying explicit negative relevance judgements is less obvious, however. Different from giving positive examples, explaining non-relevance is harder for the user [24, Section 3]. Excessive amounts of negative feedback may have negative effect on retrieval [10]. Therefore associations to the neighbours of the negative examples are not considered and $\bar{p}$ is used in the update of their $P(T)$ in (1). This scenario roughly corresponds to a nearest neighbour search, effectively eliminating seen non-marked examples from further consideration. Here $\bar{p}$ plays the role of a smoothing constant. When it equals zero, the search space is limited to the neighbours of all positive examples.

4.2 New display for the next iteration

After updating $P(T)$, a new set of objects should be presented to the user for relevance judgement. Selection of candidate objects (display update) is an important part of the search process, since it determines what the system will learn from the interaction. 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 tends to 1; or the target object(s) approach the top of the ranked list. In this article we describe experiments performed with the following three display update strategies.

4.2.1 Best-target: Following the probability ranking principle [5], $P(T)$ is considered as a score that the element receives during a retrieval session. The next display set consists of (new) objects that have largest values of $P(T)$. This 'best-target' strategy is plausible for a user unfamiliar with content-based retrieval (thus, the majority of potential users). The screen often contains objects that are the neighbours of good examples provided by the user. The user is able to observe the immediate result of his/her action. This display update strategy does not intend to explore new objects that, in the selected similarity measure, differ from the relevant ones already seen.

4.2.2 Non-deterministic strategies: In order to diversify the set of displayed elements, we introduce non-deterministic strategies as extensions to the 'best-target' one. First, we consider weighted selection of display candidates, or promotional sampling: the chance to be displayed for an object is proportional to its probability of relevance. When the distribution of $P(T)$ is peaked, proportional sampling converges to the 'best-target' method.

Second, instead of selecting the candidates with the highest score, the sample-of-best strategy makes the selection among those objects of which the probability of relevance increased since several previous iterations. This includes mainly neighbours of the relevant examples. Occasionally, elements with low but consistently growing $P(T)$ may be selected for display, giving the user a chance to see potentially relevant objects that may be very different from what he/she has already seen. Ideally, the number of elements of which $P(T)$ increases should shrink on to the group of objects that satisfy the user's information need.

5 Learning from past retrieval sessions

As stated before, there may be more than one association matrix to be used in the retrieval process. As we later show in experiments, a combination of information sources may result in better retrieval quality. Still, it is more promising to improve the existing feature estimates stored in the index.

At the end of a successful retrieval session the system is in possession of the list of objects displayed to the user $\{x_1, x_2, \ldots x_m\}$ and the corresponding relevance judgements $\{\delta_{x_1}, \delta_{x_2}, \ldots \delta_{x_m}\}$. This information can be used for improving the corresponding estimates of the probabilities $P(\delta_{x_1}|T), P(\delta_{x_2}|T), \ldots P(\delta_{x_m}|T)$.

The event of selecting x by the user as relevant/non-relevant should result in an improved estimate of the corresponding probability $P(\delta_x|T)$. To update the involved estimates, we use the maximum likelihood (ML) principle, which boils down to counting events. Let $P(\delta_x|T)$ be updated after observing one feedback action on x while seeking T. The following equation corresponds to frequency-based update for the case when δ represents binary choice:

$$P^{\text{new}}(\delta_x|T) = \frac{\kappa \cdot P^{\text{old}}(\delta_x|T) + i}{\kappa + 1}$$
$$i = \begin{cases} 1 & \text{for positive feedback} \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

Here κ is the number of observations prior to the current retrieval session. In the beginning, when there are no observations, it denotes the weight of the feature-based estimate of $P(\delta_x|T)$. Every new observation adds a unit to the denominator, and in the case of positive judgement, a unit is also added to the numerator.

Such a method of updating estimates of $P(\delta|T)$ enhances subsequent retrieval sessions. It requires an extensive interaction history, which can be achieved in, for instance, the World Wide Web environment where the number of potential users is large.

6 Related work

Maron and Kuhns in [25] talked about indexing library documents with respect to the actions taken by the searcher. The set of possible user actions includes among others (a) providing index terms for an information request, including specification of fields of interest, and (b) marking a document relevant, given the indexing terms. They suggest that this index can be refined based on the judgements of searchers. We used this work as a source of inspiration adapting it to the case when objects in the collection are atomic elements that can serve as descriptions for other members of the collection.

For modelling the interaction between the user and the system, we adapted a Bayesian framework similar to that of the PicHunter retrieval system [6]. The image relevance is determined based on the judgements of the user and (a model for) conditional dependencies between the elements. There is a large body of work dedicated to learning from relevance feedback in content-based retrieval, by far not restricted by probabilistic models. The main learning objective is to separate the answers to the query from the rest of the collection. This can be achieved in many ways: e.g. through (variations of) feature space transformation [2, 3, 26, 27], or through partitioning the data set [12, 13, 28]. The Bayesian framework turns out to be very suitable for modelling learning from the interaction, as it accommodates both short-term and long-term learning capabilities.

When using a large collection, to achieve a near-real time system response, it is not unusual that as much data as possible is processed beforehand. For example, using precomputed sets of nearest neighbours in (several) feature spaces has proved to speed up database performance during information search and browsing [29]. Strategies to combine relevance information from different sources are also studied in that work. We envision the proposed data organisation as a directed graph, but this is not unique. Stemming from a different paradigm, in [30] the collection of images is represented as a graph where images represent vertices connected to a small number of other vertices. Those are determined as nearest neighbours in numerous weighted combinations of available feature spaces. The optimal weights in a given query context is determined by the positive examples selected by the user when browsing. By quantifying in advance the feature vectors into tree-structured self-organising maps [11], relevance for images is determined by their distance on the map to the user-judged examples. Later in time, the relevance judgements make up for a separate self-organising map that is used along the feature-based maps. In [31] it is proposed to represent images with a vector of relevance judgements collected from past retrieval sessions. By applying latent semantic indexing technique, a set of closely associated images is determined and used during retrieval. This is very similar in spirit to our approach described here.

7 Experiments and evaluation

7.1 Video collection, data pre-processing and experiment set-up

The experimental evaluation is performed in the framework of 2003 TREC Video retrieval evaluation workshop (Some of the experiments performed with the collection of TREC Video-02 are presented in [32]) [22]. The video materials are CNN, ABC and C-SPAN news programs recorded between 1998 and 2001. The videos are segmented into shots, and from each shot a representative key-frame is extracted. The key-frames and shot boundaries are part of the data set, in addition to speech transcripts from a large-vocabulary automatic speech recognition system [33]. In total the test collection contains about 60 hours of recordings. Video shots represented by key-frames are the objects the system deals with. There are 24 search tasks, or topics, based on real user requests. Each search task consists of a short text description and, in most cases, few image and/or video examples of the user's information need.

7.1.1 Matrix initialisation: The experiments reported here are carried out with the following association matrices:

1. $\mathbf{M}^{t}$. Similarity between shots is computed using language models built on the text from the speech transcripts. The language model used in the experiments is described in [8].
2. Similarity between shots is based on their visual properties, namely:
   (*a*) $\mathbf{M}^{c}$. Weighted L_1 distance on three colour moments (mean, variance and skewness) in hue, saturation, value colour space of the whole image. Implemented according to [15].
   (*b*) $\mathbf{M}^{b}$. 'Bag of blocks' likelihood as described in [34]. Similarity between two key-frames (g, f) is computed as likelihood that samples taken from g could serve as a model to explain f. A draw of 100 blocks of 8×8 pixels represents the key-frame.
3. $\mathbf{M}^{t+b}$. Run-time combination of the association matrices based on textual and visual features. The combined relevance score for an object is computed as sum of the scores it achieves when using each of the matrices. The combination uses the assumption that the distribution of text-based pair-wise similarities is independent of the visual-based one. By counting neighbours in each of the two matrices to combine, we find that this assumption cannot be rejected for the text-based association matrix $\mathbf{M}^{t}$ with either of $\mathbf{M}^{c}$ and $\mathbf{M}^{b}$.

The threshold $(1-\alpha)$ is set such that on average 1.2–1.4% of possible values need to be stored. The value of $\bar{p}$ is set to 0.15 for all experiments which appears to be close to the optimum.

7.1.2 Set-up for traditional feedback method: In addition to comparison within the TREC evaluation framework, we implemented the MARS relevance feedback algorithm as described in [3]. The feature space is the one used for computation of $\mathbf{M}^{c}$, that is, three colour moments for each of hue, saturation, value channels and L_1 norm as the distance measure. The vector components are normalised, as described in the paper, so that they have equal emphasis on the resulting similarity. The initial weights for the vector components are set uniform, to be consequently updated by the intra-weight update algorithm. A random draw of six key-frames (the number of answers to the 24 topics varied from 6 to 665) from the collection served as six image queries for each topic. The results in the graphs are averaged over these six queries.

We perform an empirical study on performance differences caused by the prior distribution of the probability of relevance. In order to provide for a number of experiments a better than arbitrary estimate of the prior probability, the text from search topics serves as a text query to match against the speech transcripts. Such a scenario is quite specific to video collections, where the speech is aligned with the image and can be seen as a surrogate annotation. In an unannotated still image collection, a text query is of little use, in addition to in the case when the text query has no match. For these situations the prior probability of relevance is determined by the image queries that we used in the MARS setup.

A retrieval session starts with browsing a display set of 12 key-frames generated by the text or an image query. The key-frames are ranked by their probability of relevance. A standard TREC evaluation metric, mean average precision (MAP), is used as a measure of user's satisfaction (see [35, Appendix]). A Wilcoxon signed rank test [36] determines whether the performance figures between two methods differ significantly.

7.2 Automated experiments

In the series of experiments referred to as 'automated', the user input is replaced with relevance judgements of the TREC assessors who play the role of a 'generic user'. The experiments are carried out on a subset of the collection selected such that half of the key-frames are relevant to at least one of the 24 topics. This yields a set of 4096 shots of the 2003 collection. In this way we could test the proposed probabilistic framework, and find the best set-up to be used in the experiments with real users. A selection of automated experiments has been repeated on the whole 2003 TREC Video test data set containing about 32 000 key-frames. The results are consistent; note that mean average precision number is lower when all key-frames are used, owing to the lower proportion of relevant shots.

7.2.1 Effect of truncation on the association matrix: Mean average precision curves for $\mathbf{M}^t$ are shown in Fig. 1. Two situations, when using one of the six images as a query, and when querying with words from the TREC topics descriptions, are shown. Keeping only significant $P(\delta_x|T)$ leads to the increase of mean average precision compared to using all pairs of conditional probabilities, both with visual- and text-based matrices. This evidence confirms the choice of the α-value to determine significant pair-wise similarities (Section 3).

Comparing to a vector space model: Figure 2 shows search progress with the colour moments-based visual feature. Clearly, learning the optimal weights for the vector components does not perform better than the learning within the Bayesian framework. It is interesting to note, that on average, the set of optimal weights found by the algorithm after 20 iterations, emphasises the same vector components that have larger weights in [15], which we used in the implementation. It is worth mentioning that the vector-based model cannot straightforward take advantage of an initial text query.

7.2.2 Display update strategies: The 'best-target' display update with an *ad hoc* tuned value of $\bar{p}$ offers great improvement over iterations, both with the text and visual queries. By making sure that the user does not see the same object twice, the danger of getting stuck in an isolated island of nearest neighbours.

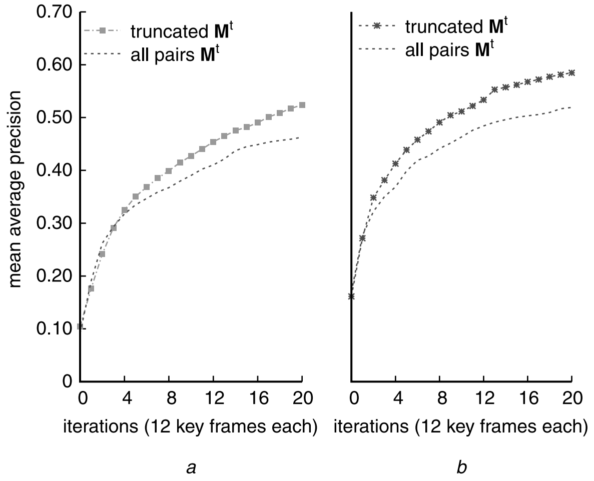

Fig. 1 *All pairs and truncated conditional probabilities for* $\mathbf{M}^t$
a Image query
b Text query

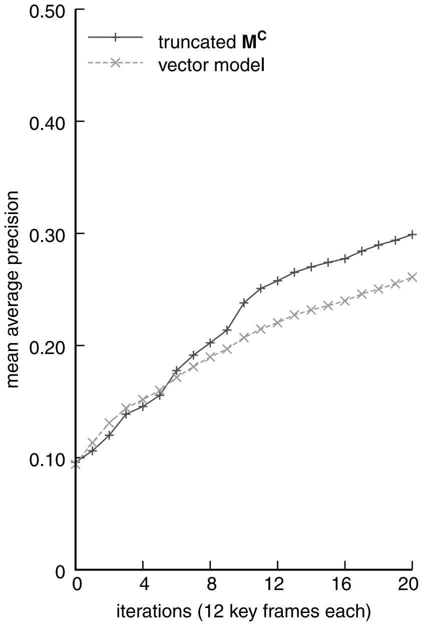

Fig. 2 *All pairs and truncated conditional probabilities using* $\mathbf{M}^b$*, and MARS relevance feedback model [3]*
Image queries used

Proportional sampling does not differ substantially from the 'best-target' strategy. When the probability of relevance distribution becomes peaked, so that few elements share most of the probability mass, the system tends to select the next display set among those few elements. Because they also occupy the top of the ranked list, the display turns out to be quite similar to that of the 'best-target' strategy (on average they share 68% of displayed objects when using text query, and 43% with image query). Consequently, there is no significant difference in mean average precision.

The 'sample-of-best' strategy in practice does not perform better than the deterministic 'best-target' method (see Fig. 3). This does not support the consideration that sampling among the promising candidates with increasing $P(T)$ can potentially give a better collection representation.

7.2.3 Combination of different modalities: Figure 4 shows mean average precision curves as a result of combining the scores from different matrices. From the graphs one can conclude that combining different sources of information improves the retrieval quality, in terms of mean average precision, by several percentage points. The combination of two matrices both based on visual appearance of the key-frames, namely M^b and M^c, does not have such effect and at best results in mean average precision that not exceeding the one that performs better. This is an expected result, since the score based on the colour statistics does not bring in new information compared to the Bag-of-Blocks likelihood score, that already has the color information in it.

7.3 Live experiments

In the live experiments, the search tasks have been performed on the 2003 TREC Video collection by real users. All of them are students from the University of Twente aged between 19 and 26. Each search task took at most 15 minutes. None of the users were familiar with the search system or were related to the development of it.

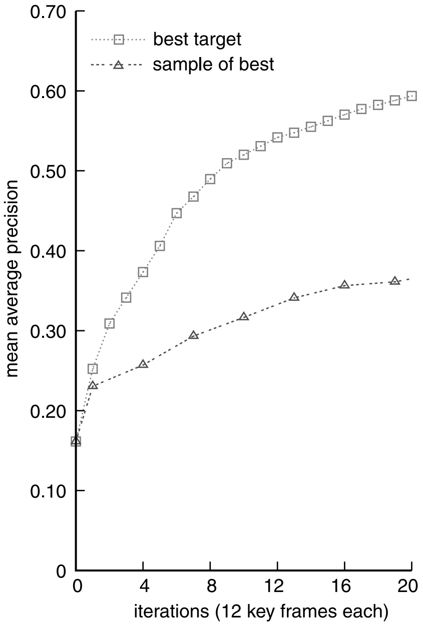

Fig. 3 *Display update strategies for $\mathbf{M}^b$ matrix, with the text query*

Proportional sampling is not plotted

A large proportion of the users' positive feedback turns out to be relevant according to the ground truth (see 'Agreement' in Table 1), thus the described automated experiments can indeed serve as an approximation to real life (see [37] for an analysis of agreement between the TREC assessors).

The set-up for live search sessions is similar to the automated experiments, using 'best-target' display update strategy, $\mathbf{M}^b$ as the access index. Words from the descriptions of search tasks served as the text query. The users retrieved key-frames (images), and not the corresponding videos. The resulting mean average precision at the end of the 'best-target' experiment is 0.245 (see Table 1), which is seventh best mean average precision for that year. For this run, 78% of the shots selected by the user were relevant according to TREC. At the same time, 48% of all relevant shots that have been displayed, were not marked as such. The users tend to miss some relevant key-frames from the display sets that contained many of those. Partially, the relevant items are missed owing to the fact that the user observed still frames, and not the video shots themselves. Therefore the key-frames of the relevant shots that do not show the required object or scene have not been marked as relevant. This issue can be resolved by enabling video display in the interface.

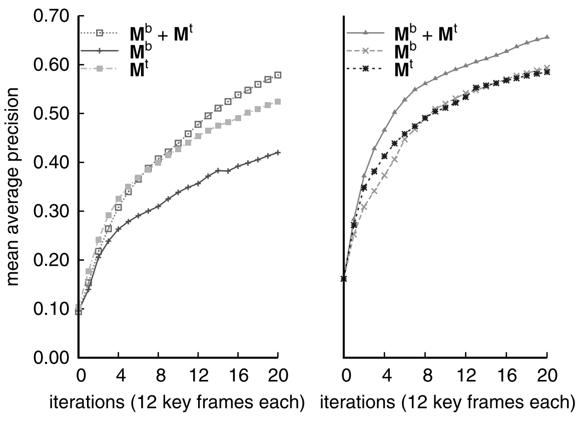

Fig. 4 *Combinations of different information sources*

a Image query
b Text query

Table 1: TRECVID experiments with real users

System type	MAP	Agreement with NIST	Missed relevant
Best-target	0.245	78.74%	48.98%
Uniform sampling	0.026	55.00%	31.25%

In the other experiment that showed the user screens sampled uniformly (mean average precision 0.026), the proportion of missed shots is much lower (31%), in addition to the lower agreement with TREC committee (55%). This is an indication that the users are inclined to mark a larger proportion of the displayed key-frames as relevant when little of those are on the screen, i.e. the independence assumption used in (1) apparently does not hold.

7.4 Learning from live experiments

We use the feedback data collected from six live retrieval sessions to conduct a preliminary experiment with training the association matrix from user feedback. To be able to use a controlled environment, these experiments are automated. However, the training data itself comes from real user sessions and contains user erroneous responses regarding the ground truth provided by TREC, mentioned in Section 7.3. The estimates of conditional probabilities for the key-frames that have been displayed to the users in the search sessions are updated using a ML method described in Section 5, equation (2). The key-frames marked as relevant serve as 'targets' in the training. After the training, the automated experiment is repeated using the same set-up, but with the new, updated association matrix, in this case M^b. As shown in Fig. 5, mean average precision is substantially increased, and the improvement is statistically significant. The most increase in mean average precision is observed at the beginning of the

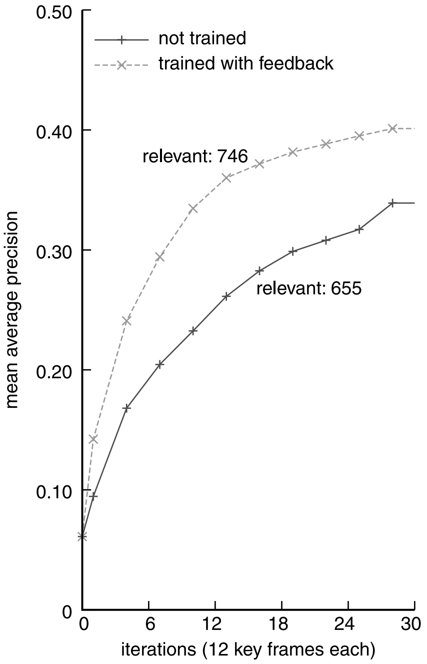

Fig. 5 *Automated experiments before and after training on real user feedback, using complete 2003 TREC Video test collection*

search, so that the user sees the desired objects earlier. The difference in mean average precision is not only due to more favourable re-arrangement of the relevant shots (mean average precision is sensitive to having relevant shots on top of the ranked list). In the experiments with the trained association matrix more relevant objects have been 'displayed' to the user (shown in the Figure).

8 Discussion and open questions

We found that the proposed image feature normalisation and smoothing by replacing similarities below a certain threshold with a constant, in the investigated visual-based and text-based feature spaces, results in higher mean average precision compared to the method that uses all pairwise similarity values. To perform the normalisation and truncation of the association matrix, we used statistics of a particular collection we wanted to search in. On a collection of video frames, our probabilistic model that uses colour moments for indexing, performs slightly better than a vector-space model using the same feature set, although we are aware of the limited nature of this comparison. The advantage of the proposed framework is that it can deploy any available technique of image understanding, to create an initial association matrix.

Truncation of the matrix enables efficient combination of different similarity measures, such as visual information from key-frames and transcripts of the speech occurring in video shots. Combination of independent sources of information has a positive effect on the retrieval. We used equal weights when adding up the scores, but there should possibly be better weighting schemes—this needs to be investigated.

So far we did not observe any improvement when attempting to efficiently diversify the set of displayed objects as compared to the common strategy of showing the meet relevant candidates. Although, from the point of view of optimal learning, the 'best-target' is not the best choice, it is hard to compete with when real-life data is used in the collection.

Keeping only the significant values allows an interactive retrieval system to ensure fast response time, which is a necessary condition for an interactive retrieval system. This in turn will provide vast amount of training data. Learning from the history of relevance judgements in retrieval sessions with the real users substantially improves the successive searches. The improvement is observed not only due to higher ranking of the previous positive examples, but also due to a larger number of relevant key-frames that are displayed to the user. More advanced learning techniques are needed to update the conditional probability estimates between the unseen objects.

9 Acknowledgment

The authors would like to thank Thijs Westerveld of CWI Amsterdam for providing BoB likelihood scores for this collection.

10 References

1 Flickner, M., Sawhney, H., Niblack, W., and Ashley, J.: 'Query by image and video content: the QBIC system', *Computer*, 1995, **28**, (9), pp. 310–15

2 Ishikawa, Y., Subramanya, R., and Faloutsos, C.: 'MindReader: querying databases through multiple examples'. Proc. 24th Int. Conf. Very Large Data Bases, VLDB, 1998, pp. 218–227

3 Rui, Y., Huang, T., Mehrotra, S., and Ortega, M.: 'A relevance feedback architecture in content-based multimedia information retrieval systems'. Proc. IEEE Workshop on Content-based Access of Image and Video Libraries, in conjunction with IEEE CVPR, 1997

4 Hiemstra, D.: 'Using language models for information retrieval'. PhD thesis, University of Twente, 2001

5 Robertson, S.E.: 'The probability ranking principle in IR', *J. Doc.*, 1977, **33**, (4), pp. 294–304

6 Cox, I.J., Miller, M.L., Minka, T.P., Papathomas, T.V., and Yianilos, P.N.: 'The Bayesian image retrieval system, PicHunter: theory, implementation, and psychophysical experiments', *IEEE Trans. Image Process.*, 2000, **9**, (1), pp. 20–37

7 Vasconcelos, N.M.: 'Bayesian models for visual information retrieval'. PhD thesis, Massachusetts Institute of Technology 2000

8 Westerveld, T., de Vries, A.P., van Ballegooij, A.R., de Jong, F.M.G., and Hiemstra, D.: 'A probabilistic multimedia retrieval model and its evaluation', *EURASIP J. Appl. Signal Process.*, 2003, **2003**, (2), pp. 186–198

9 Rocchio, J.J.: 'Relevance feedback in information retrieval' in Salton, G. (Eds): 'The smart retrieval system: experiments in automatic document processing' (Prentice Hall, 1971), pp. 313–323

10 Müller, H., Müller, W., Marchand-Maillet, S., Pun, T., and Squire, D.: 'Strategies for positive and negative relevance feedback in image retrieval'. Proc. Int. Conf. on Pattern Recognition, Barcelona, Spain, 2000

11 Laaksonen, J., Koskela, M., Laakso, S., and Oja, E.: 'Self-organizing maps as a relevance feedback technique in contentbased image retrieval', *Pattern Anal. Appl.*, 2001, **4**, (2-3), pp. 140–152

12 Drucker, H., Shahrary, B., and Gibbon, D.C.: 'Relevance feedback using support vector machines'. Proc. 18th Int. Conf. on Machine Learning, Morgan Kaufmann, San Francisco, CA, 2001, pp. 122–129

13 Tong, S., and Chang, E.: 'Support vector machine active learning for image retrieval'. Proc. Ninth ACM Int. Conf. on Multimedia, ACM Press, 2001, pp. 107–118

14 Tamura, H., Mori, S., and Yamawaki, T.: 'Texture features corresponding to visual perception', *IEEE Trans. Syst. Man Cybern.*, 1978, **6**, (8), pp. 460–473

15 Stricker, M.A., and Orengo, M.: 'Similarity of color images'. Storage and Retrieval for Image and Video Databases (SPIE), San Diego/La Jolla, California, USA, 1995, pp. 381–392

16 Pentland, A., Picard, R., and Sclaroff, S.: 'Photobook: content-based manipulation of image databases', *Int. J. Comput. Vis.*, 1996, **18**, (3), pp. 233–254

17 Gudivada, V.N., and Raghavan, V.V.: 'Design and evaluation of algorithms for image retrieval by spatial similarity', *ACM Trans. Inf. Syst.*, 1995, **13**, (2), pp. 115–144

18 Faloutsos, C.: 'Searching multimedia databases by content' (Kluwer Academic Publishers, Boston, USA, 1996)

19 Weber, R., Schek, H.J., and Blott, S.: 'A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces'. Proc. 24th Int. Conf. on Very Large Data Bases(Morgan Kaufmann Publishers Inc., 1998, pp. 194–205

20 Faloutsos, C., and Lin, K.-I.: 'FastMap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets'. in Carey, M.J., and Schneider, D.A. (Eds): Proc. 1995 ACM SIGMOD Int. Conf. on Management of Data, (ACM Press), 1995, pp. 163–174

21 De Vries, A.P., Mamoulis, N., Nes, N., and Kersten, M.L.: 'Efficient k-NN search on vertically decomposed data'. ACM SIGMOD Int. Conf. on Management of Data, Madison, WI, USA, June 2002

22 Smeaton, A., Kraaij, W., and Over, P.: 'TRECVID 2003 - an introduction'. Text Retrieval Conf. TRECVID Workshop, National Institute of Standards and Technology, 2003, URL http://www-nlpir.nist.gov/projects/tv2003/tv2003.html

23 Aksoy, S., and Haralick, R.M.: 'Feature normalization and likelihood-based similarity measures for image retrieval', *Pattern Recognit. Lett.*, 2001, **22**, (5), pp. 563–582

24 Ruthven, I., and Lalmas, M.: 'A survey on the use of relevance feedback for information access systems', *Know. Eng. Rev.*, 2003, **18**, (2), pp. 95–145

25 Maron, M.E., and Kuhns, J.L.: 'On relevance, probabilistic indexing and information retrieval', *J. Assoc. Comput. Mach.*, 1960, **7**, pp. 216–244

26 Santini, S., and Jain, R.: 'Integrated browsing and querying for image databases', *IEEE Multimedia*, 2000, **7**, (3), pp. 26–39

27 Rui, Y., and Huang, T.: 'Optimizing learning in image retrieval'. Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, 2000

28 MacArthur, S., Brodley, C., and Shyu, C.: 'Relevance feedback decision trees in content-based image retrieval'. Proc. IEEE Workshop on Content-based Access of Image and Video Libraries (CBAIVL'00), South Carolina, USA, 2000

29 Fagin, R.: 'Combining fuzzy information from multiple systems', *J. Comput. Syst. Sci.*, 1999, **58**, pp. 83–99

30 Heesch, D., and Rüger, S.M.: 'NN^k networks for content-based image retrieval', *Lect. Notes Comput. Sci.*, 2004, **2997**, pp. 253–266

31 Heisterkamp, D.R.: 'Building a latent semantic index of an image database from patterns of relevance feedback'. Proc. 16th Int. Conf. on Pattern Recognition (ICPR 2002), Quebec, Canada, 2002, **4**, pp. 134–145

32 Boldareva, L., and Hiemstra, D.: 'Interactive content-based retrieval using pre-computed object-object similarities'. Proc. Int. Conf. on

Image and Video Retrieval, CIVR, Dublin, Ireland, July, 2004, pp. 308–316

33 Gauvain, J.L., Lamel, L., and Adda, G.: 'The LIMSI broadcast news transcription system', *Speech Commun.*, 2002, **37**, (1-2), pp. 89–108

34 Westerveld, T., Ianeva, T., Boldareva, L., de Vries, A.P., and Hiemstra, D.: 'Combining information sources for video retrieval'. Text Retrieval Conf. TRECVID Workshop, National Institute of Standards and Technology, Gaithersburg, MD, USA, 2003

35 In: Voorhees, E.M., and Harman, D.K., eds, The Tenth Text Retrieval Conference (TREC-2001)(Department of Commerce, National Institute of Standards and Technology, Gaithersburg, MD, USA, 2002

36 Ott, L.: 'An introduction to statistical methods and data analysis', (Boston: PWS-KENT, 1988, 3rd edn.)

37 Voorhees, E.M.: 'Variations in relevance judgments and the measurement of retrieval effectiveness', *Inf. Process. Manage.*, 2000, **36**, (5), pp. 697–716