Jointly Optimising Relevance and Diversity in Image Retrieval

Thomas Deselaers[1,2]
[1]Computer Vision Laboratory
ETH Zurich
Zurich, Switzerland
deselaers@vision.ee.ethz.ch

Tobias Gass[2], Philippe Dreuw[2], Hermann Ney[2]
[2]RWTH Aachen University
Computer Science Department
Aachen, Germany
{gass,dreuw,ney}@cs.rwth-aachen.de

ABSTRACT

In this paper we present a method to jointly optimise the *relevance* and the *diversity* of the results in image retrieval. Without considering diversity, image retrieval systems often mainly find a set of very similar results, so called near duplicates, which is often not the desired behaviour. From the user perspective, the ideal result consists of documents which are not only relevant but ideally also diverse. Most approaches addressing diversity in image or information retrieval use a two-step approach where in a first step a set of potentially relevant images is determined and in a second step these images are reranked to be diverse among the first positions. In contrast to these approaches, our method addresses the problem directly and jointly optimises the diversity and the relevance of the images in the retrieval ranking using techniques inspired by dynamic programming algorithms. We quantitatively evaluate our method on the ImageCLEF 2008 photo retrieval data and obtain results which outperform the state of the art. Additionally, we perform a qualitative evaluation on a new product search task and it is observed that the diverse results are more attractive to an average user.

Categories and Subject Descriptors

H.3 [**INFORMATION STORAGE AND RETRIEVAL**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation, Performance

Keywords

Image Retrieval, Similarity, Diversity, Clustering

1. INTRODUCTION

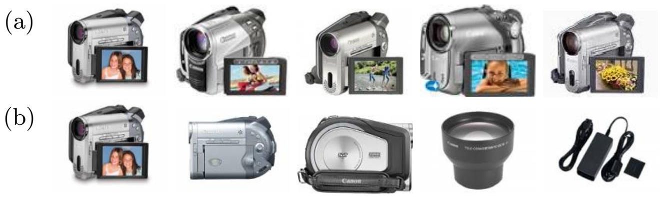


Figure 1: Example for a similarity-based image search (a) and the desired diversity-enhanced image search (b).

Image retrieval in general and content-based image retrieval in specific are a topic that has been investigated quite thoroughly over the past years and considerable progress has been achieved in searching and finding similar images for a given query image [8, 23, 24]. In particular similarity-based image search is able to find very similar images, so called *near duplicates* and commonly it is assumed that similar images are relevant to a given query image. For a user of an image retrieval system, this is not necessarily the desired output of a system, instead it might rather be desired to obtain different images which only share certain properties of the given query image. In Fig. 1 we give two example results for a textual query with the exact name and make of a certain camcorder. Fig. 1(a) is a typical example of too homogeneous results and Fig. 1(b) shows an example of a diverse set of results, depicting the object of interest and certain accessories that might be interesting to the user. Such results are not necessarily possible to obtain using only visual methods. In our approach, we fuse visual and textual cues, which convey different information, in order to obtain such results. That is, we still assume that relevant images are similar (either according to the textual annotation or visually), but we want to avoid returning all near duplicate images and instead try to find a set of results which are as diverse as possible but still relevant with respect to the query.

A similar effect was observed in the literature on recommendation engines, where accuracy is not the only criterion to satisfy a user but additionally users want *diverse* recommendations [13].

A typical application where diverse results in information retrieval are desirable is the retrieval of product images. E.g. a user interested to buy a new cell phone queries his favourite shopping portal with the name and make of his current cell phone in order to find the successor model. The search functionality of the site then commonly delivers many cell-phones from the same brand but alternatively it could suggest models from different makers which share some of the features. Other examples might be to start with the frontal view of an object and find different view points or distinguish between official product images and images that

CIVR '09, July 8-10, 2009 Santorini, GR
Copyright 2009 ACM 978-1-60558-480-5/09/07 ...$5.00.

amateurs put online, e.g. in online auction websites.

Ideally, these approaches would be combined with an effective image browsing tool to allow for fast navigation in the space of images. User interfaces that combine similarity search with visualisation of image similarity in a 3D space are presented in [14, 15] and allow for effective searching and browsing of large collections without explicitly considering diversity.

Diversity of the results was addressed in the ImageCLEF 2008 photo retrieval evaluation [2] where additionally to system accuracy the diversity of the retrieval results was evaluated. In this evaluation, 24 groups participated and submitted slightly more than 1,000 runs, where most of the runs were a combination of textual and visual information retrieval. Nearly three quarters of the participating groups used techniques to explicitly improve the diversity of their results. In all presented approaches this was achieved using a two-step strategy: first, a normal retrieval was performed and second, the results were reranked using a clustering step of the N-best results. The methods presented here address the two contradicting objectives of similarity/relevance and diversity jointly.

The contributions of this paper are four-fold:

- We analyse and discuss the current state of the art in diversity-aware image and information retrieval.
- We propose a criterion to assess the diversity of a ranked list which can be combined with a common image/information retrieval criterion. This joint criterion is not limited to a particular domain but can be used in any image or information retrieval context where it is possible to define a similarity or dissimilarity measure for a pair of images or documents.
- We present three algorithms using this criterion for retrieval: a greedy approximation, an approximation inspired by dynamic programming (DP) techniques, and a DP algorithm which is able to find an optimal solution under a simplified diversity criterion.
- We compare the proposed techniques experimentally to a clustering-based diversification technique on two image retrieval tasks. We quantitatively evaluate our methods on the ImageCLEF 2008 photo retrieval task and show preliminary results on a novel product search task, which both incorporate visual and textual information.

The remainder of this paper is structured as follows: in Section 2 we discuss the related work. In Section 3 we present a criterion to measure the diversity of a list of images or documents. In Section 4 we present three approaches to obtain a diverse list of results which are experimentally evaluated in Section 5 and in Section 6 the paper is concluded.

2. RELATED WORK

As described above, the problem of too homogeneous results has recently been observed in the literature on recommendation engines [13].

This problem has been analysed in more detail in several other papers. McGinty and Smyth [12] investigate the role of diversity in recommender systems and propose an iterative procedure to improve the user experience when interacting with the system. They conclude that the user will find satisfactory result much faster if the diversity of the results is higher. However, they also note that too high diversity has the risk of losing relevant items and that thus a trade off is required. Similarly, Xu and Yin [25] compare different information retrieval approaches with respect to the relevance and the diversity of the retrieved results and conclude that the user-centred information retrieval community should join forces with the system-centred information retrieval community to obtain optimal performance and user satisfaction. Fleder and Hosanagar [6] hypothesize that diverse results can have an impact on the sales of certain products and evaluate their ideas in simulated experiments. Diversity in social networks has been investigated by Lemire et al. [10].

Also in the domain of recommendation engines some approaches to enhance the diversity of the results were proposed. Ziegler et al. [28] propose a technique to diversify top-N lists of recommendation engines and observe in a user study that user satisfaction is improved even though the average accuracy of the lists is deteriorated. An interesting approach was presented by Le and Smola [9]. Instead of addressing the problem of diversity directly, they propose a ranking technique that can be optimised with respect to nearly arbitrary ranking measures and suggest an extension of their approach to incorporate diversity constraints but do not discuss a solution to this.

The paper which is most closely related to our approach was presented by Zhang and Hurley [27]. Similar to our approach, they pose the problem of finding relevant and diverse results as a joint optimisation problem and propose three different approaches using binary linear and binary quadratic programming techniques respectively. In contrast to our approach they do not create ranked lists of results. Their method delivers result sets without any particular ordering. Chen and Karger [3] present a Bayesian retrieval approach that also incorporates diversity in the retrieval criterion and propose a greedy approximation for retrieval.

Zhai et al. [26] and Clarke et al [4] discuss various measures to evaluate the diversity of the delivered results in an information retrieval system under the name of subtopicality. They propose cluster recall as a suitable measure which was also used in the ImageCLEF 2008 task [2] and which we also use for evaluation.

In image retrieval, the ImageCLEF 2008 photo retrieval task has made many groups aware of the issue of diversity in information retrieval and several groups have tried to address this problem by first retrieving a list of candidate images and then reranking this list using a clustering algorithm in a post-processing step, e.g. [5, 18, 22].

3. DIVERSITY

Although for a human it is pretty easy to judge whether a list of images is diverse or not, so far no formal definition has been proposed. Similar to [27], we base our measure on the dissimilarity between the individual items considered.

Let q be a query, $\mathcal{B} = \{x_1, \ldots, x_N\}$ be the database of possible documents. For a given partial result $(x_{r_1}, \ldots, x_{r_J})$ we define the *novelty* for a candidate image x^* to be added to the list at position $J + 1$ as

$$\mathcal{N}(x^*; (x_{r_1}, \ldots, x_{r_J})) = \frac{1}{J} \sum_{j=1}^{J} d(x^*, x_{r_j}), \qquad (1)$$

where $d(x^*, x_{r_j})$ is a normalised dissimilarity measure between the candidate and the individual images in the results. That is, we give a high novelty score to those images which are on average dissimilar to the current set of results.

Instead of considering the average dissimilarity of the hypothesised image it is possible to consider the dissimilarity to the most similar image as a novelty score which can be obtained from Eq. (1) by replacing the sum with a minimum-operator.

Considering the similarity retrieval score for an image which is defined as

$$\mathcal{S}(x^*; q) = 1 - d(x^*, q) \tag{2}$$

it can be observed that these two measures are somewhat controversial because images that are very similar to the query are also likely similar among each other and thus have a low novelty score.

To fuse these two criteria into a retrieval score $\mathcal{R}$ for a candidate image x^* given a query q and a set of hypothesised prior results we use a weighted sum

$$\begin{aligned}\mathcal{R}(x^*; q, (x_{r_1}, \dots, x_{r_J}) &= \alpha\mathcal{S}(x^*; q) \\ &+(1.0 - \alpha)\mathcal{N}(x^*; (x_{r_1}, \dots, x_{r_J}))\end{aligned} \tag{3}$$

where a higher α means that similarity to the query is more important than diversity, and a low α means that diversity is more important than similarity. It can be expected that $\alpha < 0.5$ will not yield good results since even a user who is looking for diverse results still expects the results to be relevant [12].

Now we can combine the proposed measures to be able to judge the similarity and the diversity of a ranked list by summing over the individual images

$$\begin{aligned}&\mathcal{F}(q; (x_{r_1}, \dots, x_{r_J})) \\ &= \sum_{j=1}^{J} \mathcal{R}(x_{r_j}; q, (x_{r_1}, \dots, x_{r_{j-1}}), \\ &= \sum_{j=1}^{J}\left(\alpha\Big[1 - d(x_{r_j}, q)\Big] + (1-\alpha)\Big[\frac{1}{j}\sum_{j'=1}^{j} d(x_{r_j}, x_{r_{j'}})\Big]\right)\end{aligned} \tag{4}$$

and by finding the list of indices $(r_1, \dots, r_J)$ which maximises this score, we can obtain a retrieval result which satisfies similarity as well as diversity requirements:

$$(\hat{r_1}, \dots, \hat{r_J}) = \arg\max_{(r_1,\dots,r_J)} \left\{\mathcal{F}(q; (x_{r_1}, \dots, x_{r_J}))\right\}. \tag{5}$$

Note that for the evaluation of this criterion no properties of the feature vectors are required except that it has to be possible to calculate the scores from Eqs. (1)-(3) which only require the possibility to compute dissimilarity measures between a pair of images.

4. TECHNIQUES TO OBTAIN DIVERSE BUT RELEVANT RESULTS

We introduce three algorithms to obtain diverse but relevant results in image and information retrieval. The first method represents the multi-step clustering-based state of the art and thus does not consider the diversity criterion presented above. The other approaches directly optimise the diversity criterion presented above.

4.1 Clustering as Post-Processing

As described above, most approaches to obtain improved diversity of the retrieval result use a clustering step, commonly k-means with a fixed number of clusters k. To create the final ranked list an image is taken from each cluster in a round-robin fashion and added to the result [5, 18, 22].

In our implementation, we use the LBG clustering algorithm [11] which is an extension of the EM algorithm for Gaussian mixtures. It starts with a single density which is then incrementally split until the desired number of clusters is reached. Note that this clustering technique requires to be able to compute means of the cluster prototypes and is therefore not suitable to be used with arbitrary dissimilarity measures in the retrieval process. We are aware that clustering techniques that overcome this problem, such as spectral clustering [19], exist but decided to use this technique as it is known to be a solid baseline clustering technique and in contrast to spectral clustering, which commonly creates only two clusters, allows to obtain an arbitrary number of clusters.

Then, the retrieval consists of three steps: In the first step, we rank all images according to their similarity scores $\mathcal{S}$ from Eq. (2). In the second step, we take the first $n \leq N$ images of the ranked list and cluster them using our clustering algorithm. In the last step, the outcome of the clustering is used in order increase the diversity of the top 20 results. That is, we assume that the individual clusters represent different subtopics or clusters of results. Therefore we want to have the top results chosen to represent as many of the clusters as possible. In order to achieve this, we change the scores for the individual images x after J images have been ranked by adding a weighted novelty bonus

$$b(x) = 1 - \frac{\sum_{j=1}^{J} \delta(c(x_j), c(x))}{CJ}, \tag{6}$$

where C is the number of clusters, J is the number of images already in the ranked list, and $c(x)$ is the cluster for image x. Thus, this novelty bonus is highest for the images from the cluster which is least represented in the previously returned results.

4.2 Greedy Selection

In the *greedy* algorithm, we consider the retrieval score $\mathcal{R}$ as defined in Eq. (3) and incrementally add the image with the highest score.

Thus, first the image which is most similar to the query is added to the result list, because there are no results yet and thus the similarity dominates the term. Then, for each of the remaining images x we determine $\mathcal{R}(x; q, (x_{r_1}, \dots, x_{r_J})$ and add the image with the highest values to the list until all images are ranked.

A problem with this method is that the scores for the images change with each image that is added to the list of results and thus a disadvantageous choice at an early position in the result may significantly deteriorate the entire ranking.

4.3 Dynamic Programming

In this approach, we leverage the issue of non-optimal rankings from the greedy algorithm by optimising the ranking using an algorithm which is inspired by DP which does not consider only one path but which considers many paths simultaneously. The number of total rankings is exponential and thus it is infeasible to evaluate all paths. The problem, has certain similarities to the travelling salesman problem

(TSP) because we are looking for an optimal path in a fully connected graph where each vertice (image) has to be visited (ranked) exactly once. It differs to the TSP because the weights (scores, in particular the novelty score $\mathcal{N}$) for the individual edges change depending on the images already ranked, and thus in order to find the *optimal* ranking it is necessary to evaluate $N!$ rankings. Here we use an algorithm which is inspired by DP and evaluates only $N \times N$ paths but was shown to perform well for many instances of the TSP [17].

The optimisation is done starting with the first image to be returned and then incrementally for each position in the ranked list every image is hypothesised considering every other image as predecessor. In this step, for each image at each position in the list, the list of all hypothesised predecessors is memorised which allows for an efficient calculation of the novelty score $\mathcal{N}$ from Eq. (1).

This algorithm works similar to common DP methods, i.e., we define an auxiliary function $Q(j, x)$ which denotes the score for the best ranked list of length j where image x is at position j. Then, we can define the recursive equation for the DP problem

$$\begin{aligned} &Q(j,x) \qquad (7)\\ &= \max_{x'} \left\{ Q(j-1,x') + \mathcal{R}\left(x; q, (x_{r_1}, \ldots, x_{r_{j-2}}, x')\right) \right\} \end{aligned}$$

which considers every image x' to be ranked on position $j-1$ building on the solution corresponding to $Q(j-1, x')$. In order to allow for reconstructing the best solution among the evaluated ones, commonly a backpointer array is created which for the problem at hand would be defined as follows:

$$\begin{aligned} &B(j,x) \qquad (8)\\ &= \arg\max_{x'} \left\{ Q(j-1,x') + \mathcal{R}\left(x; q, (x_{r_1}, \ldots, x_{r_{j-2}}, x')\right) \right\}. \end{aligned}$$

However, since the calculation of $\mathcal{F}$ depends on the hypothesised solution up to that point, it would be necessary to perform the trace-back in each step. In order to avoid this, we reorganise the trace-back structure and store the full list of predecessors as backpointers in each step. To reduce the memory usage of this approach, we do not keep the full array of backpointers for each position but since the propagation of full trace-backs spares us the necessity to trace-back the solution at the end of the algorithm, we only need two columns of the trace-back, one for the ranking position that is currently being processed and one for its predecessor. Once a column is completed, the trace-back of the predecessor can be discarded since it is a subset of the current one.

In order to avoid having duplicate images in the results, we set the novelty $\mathcal{N}$ to $-\infty$ for every image that is already part of the hypothesised result list .

A desired result for this algorithm is of course that the most similar image, say $\hat{x}$ is ranked first. This is obviously the desired choice since it has the highest similarity and it has the same novelty as any other image which is returned in this position. However, due to the non-symmetric definition of novelty which is only looking backwards, it is not always the choice of the algorithm. Therefore, instead of setting the boundary conditions $Q(0, x)$ to 0 for every x, we set $Q(0, \hat{x}) = \mathcal{S}(\hat{x}, q)$ and $Q(0, x) = 0$ for all other images. This initialisation guarantees that the most similar image is always returned first.

In order to rank a database of N images, this algorithm has a complexity of $\mathcal{O}(N^4)$, because we need N steps, where in each step, all N images and all N possible direct predecessors have to be hypothesised and furthermore, in the computation of the novelty score $\mathcal{N}$ all predecessors of the ranked list have to be considered, which are maximally N many. However, all computations involved are cheap since it is possible to precalculate the required dissimilarity measures between the images.

Furthermore, a significant speedup can be obtained if not the full database is ranked but only the top M images from a normal retrieval step on the database. In informal experiments, we have investigated how many images have to be considered in order to get an identical ranking in the top 20 results and found that for $M = 100$ no difference in the top-20-ranking was observed.

4.3.1 DP for a Simplified Problem

As described above, a problem with this approach is that we cannot evaluate all possible rankings efficiently due to the dependency of the novelty score on all predecessors. The algorithm proposed above therefore only considers a subset of all possible solutions. Another option is to change the objective function to allow for an exhaustive search of the hypothesis space. Here, the problem is rewritten as searching for the optimal subsequence of the initial ranking with respect to a given criterion. Therefore, we modify our objective function in two respects:

- The novelty function (Eq. (1)) is changed to consider only the direct predecessor
$$\mathcal{N}_s(x^*; (x_{r_1}, \ldots, x_{r_J})) = d(x^*, x_{r_J}). \qquad (9)$$
- The desired result is a monotonous subsequence of the initial similarity ranking that maximises the criterion (4) with the modified novelty score (9).

With these two assumptions, it is possible to apply a conventional, non-approximative DP algorithm to find the optimal (monotonous) subsequence $(x_{r_1}, \ldots, x_{r_J})$ from the sorted sequence of similar images $(x_1, \ldots, x_N)$, where $\{r_1, \ldots, r_J\} \subset \{1, \ldots, N\}$ and $r_i < r_{i+1}$.

Comparing the two approaches based on DP, it can be observed that the first one tries to optimise the desired criterion but is not guaranteed to find the optimal solution due to search errors, whereas the second algorithm optimises a simplified criterion for which it will definitely find the optimal solution.

5. EVALUATION

We compare the proposed methods quantitatively on the ImageCLEF 2008 photo database and show some additional qualitative results on a novel product image dataset.

IAPR TC-12/ImageCLEF 2008 photo retrieval database.

The ImageCLEF 2008 photo retrieval task [2] built on the IAPR TC 12 database [7] which is available online[1]. This dataset consists of 20,000 images where each image is annotated in English, German, and Spanish. Additionally, the ImageCLEF team defined 60 queries which were used in 2006 and 2007. In 2008, 39 of these topics were reused. For these 39 queries, different clusters in the results were annotated

[1] http://www.imageclef.org/photodata

Figure 2: Example images from (a) the ImageCLEF 2008 photo retrieval task and (b) the product images database.

Table 1: Product database statistics.

Task	#concepts	#images	avg./concept
Camcorders	78	6,768	86.76
Cameras	90	4,747	52.74
Cellphones	62	11,650	187.90
Laptops	57	1,834	32.17
MP3-Players	88	2,281	25.92
Multimedia Players	39	2,852	73.12
TV	216	3,609	16.70
Total	630	33,741	53.55

manually to allow for measuring the diversity (according to these clusters) of the delivered results. Four example images from this dataset are shown in Fig. 2(a).

Product Database.

We use a novel product database, which was provided by Exalead. It contains seven retrieval tasks dedicated to different types of high tech products, such as camcorders, laptops, or TVs, and consists of several concepts (see Table 1). Each concept corresponds to a textual search query and consists of a set of images along with their surrounding text and the original URLs of the images.

The aim in this task is to create an ordering of the images of each concept which has diverse and relevant images ranked top (cf. Fig. 1). Example images from this dataset for Cellphones, Laptops, MP3-Players, and TVs are shown in Fig. 2(b).

System Setup.

For the experiments, the system was setup to use a combination of five different descriptors:

Colour histograms describe the distribution of colours in the images and have been shown to be an important cue in generic image retrieval applications.

Tamura texture features are a manually designed texture descriptor which have been used frequently in the literature [21].

GIST descriptors are small descriptors that capture the overall shapes of a scene [16].

Bag-of-Visual-Word Histograms were originally developed in the texture analysis domain but have been proved to be very useful for object recognition and image retrieval [20].

English Text: in the experiments with the clustering approach the annotations are represented as 2,000-bin histograms of the most frequent words in order to allow for clustering of these feature vectors. In the distance-based retrieval approaches we use a SMART-2 text retriever to deliver distances for the retrieval process.

5.1 Quantitative Evaluation on the ImageCLEF task

For evaluation we follow the protocol of the ImageCLEF 2008 photo retrieval task [2]. I.e. the database consists of 20,000 images and we process 39 queries. The assessment of the performance of the runs was done mainly with three measures: P_{20}, precision after 20 results, which best captures the performance of image search engines such as Google images which present approximately twenty images on the first page. CR_{20}, cluster recall [2, 4] after 20 results, which measures the number of subtopics/content-clusters in the first 20 results, and the combined F-measure of these two:

$$F = \frac{2P_{20}CR_{20}}{P_{20} + CR_{20}}. \tag{10}$$

For the experiments on the ImageCLEF task, we performed two different sets of experiments. A baseline setup, where the textual information is represented as histograms of the most common 2,000 words, and an improved setup, where the text is incorporated using a variant of the SMART-2 retrieval metric. The baseline setup allows to apply the clustering (cf. Section 4.1), whereas the text retrieval setup is far more powerful than the simple histogram representation and thus achieves higher performances, but the clustering approach is not applicable.

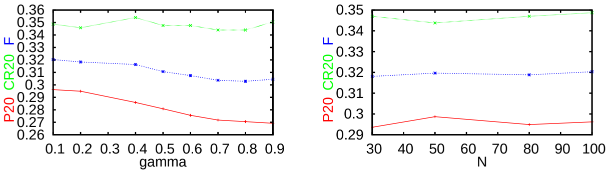
5.2 Baseline Setup

For this setup we tuned the weights for the individual descriptors on the queries from the ImageCLEF 2007 task which were not reused in the 2008 evaluation without considering diversity, and found that equal weights for all descriptors but a 3 times higher weight for the bag of visual words descriptor led to the best results.

Clustering.

Fig. 3 gives results for the evaluation of different parameters for the clustering approach. The red lines denote P_{20}, the green lines denote CR_{20}, and the blue lines denote the joint F-measure. Fig. 3(a) shows the impact of the weighting parameter γ on the results of the clustering approach. A higher γ denotes a higher weighting of the novelty bonus in the clustering result and it can be observed that with too high γ the impact of the clustering is too big and thus the images are mainly returned from unseen clusters without considering the similarity. With γ in a reasonable range, it only has a minor impact. Fig. 3(b) shows the impact of the number of images considered for clustering and this also only has a minor impact on the results. In general, the clustering approach only leads to minor improvement compared with the baseline result where no diversity-improvement is applied (cf. Table 2).

We also applied the direct approaches, greedy, DP, and monotone DP on the same setup. Results from these are given in Fig. 4. Here a low value for α is the weighting factor from Eq. (3). A high value denotes that similarity has a higher weight than diversity, i.e. if $\alpha = 1.0$ is chosen, diversity is not considered and if $\alpha = 0.0$ is chosen similarity is not considered. In Fig 4 (a)-(c), it can be observed that for all approaches a too low α leads to strongly deteriorated performance, in particular on the precision (Fig. 4 (a)), but



(a) (b)

Figure 3: The effect of the different parameters on the results of the clustering approach: (a) weight γ, (b) number of images considered. Solid lines show P_{20}, dotted lines CR_{20} and dashed lines represent the F value.

Table 2: Results from the experiments with the textual information represented as histograms

method	P_{20}	CR_{20}	F
no diversity	0.295	0.347	0.319
clustering	0.296	0.349	0.320
greedy	0.292	0.352	0.319
DP	0.295	0.354	0.322
monotone DP	0.290	0.357	0.320

also on the cluster recall (Fig. 4 (b)), and thus also on the F-measure (Fig. 4 (c)). This effect is strongest in the greedy approach (red solid lines). In Fig. 4 (d)-(f), the effect of changing the number of images considered for the optimisation is investigated. It should be noted that the monotone DP and the greedy approach are very sensitive with respect to this parameter and using too many images leads to bad performance since these algorithms then tend to prefer the dissimilar and thus possibly irrelevant images. Note, that it is also possible to consider all images at once, but as can be seen from these plots, this is not beneficial on the results.

Table 2 gives an overview of the different approaches with tuned parameters on the baseline setup. It can be observed that all approaches lead to an improvement of the diversity. The best F-measure is obtained by the DP approach, the greedy approach, the monotone DP approach, and the clustering approach have approximately the same performance. Additionally, the DP approach has the advantage to be relatively robust with respect to its parameters.

5.3 Improved Setup

In this section, we describe experiments where the textual information is incorporated using the SMART-2 retrieval metric. Thus, the clustering approach is not applicable. We tuned the weights in the same way as described above and found that bag of visual words and gist descriptors have slightly higher weights than colour and Tamura histograms, and that the textual information is weighted three times higher than the other features. This setup leads to a big improvement of the results when no diversity-increasing technique is applied and P_{20} rises from 29% to 49% (cf. Table 3).

Analogously to the experiments described above we evaluate the impact of the different parameters on the results for the three approaches that directly optimise the diversity. Fig. 5 (a) -(c) shows the impact of the α parameter on the results. Again it can be observed that for all approaches a too small α, i.e. too little weighting to the similarity, leads to bad results because the diversity is weighted so high that the returned images, albeit diverse, are not relevant anymore, which can be seen from the drop of P_{20} in Fig. 5 (d). This effect is strongest in the greedy approach (red solid line) and also relatively strong in the monotone DP approach (dotted blue line). In the DP approach (dashed green line) this effect is smoothed as the all images from the result set are considered for the diversity.

Table 3: Results from the experiments with the SMART-2 retrieval for the textual information

method	P_{20}	CR_{20}	F
no diversity	0.491	0.482	0.486
greedy	0.494	0.499	0.496
DP	0.499	0.514	0.506
monotone DP	0.483	0.484	0.484
ParisTech [5]*	0.689	0.680	0.684
XRCE [1]	0.512	0.426	0.465

* result obtained with manual tuning and user interaction.

In Fig. 5 (d)-(f) the impact of the number of images considered in the optimisation is evaluated. It can be observed that the greedy approach (solid red line) performs best if only very few images are available, which can probably be explained by the fact that the top few images are all relevant, and this approach prefers extremely dissimilar and thus likely irrelevant images if it has the choice. The monotone DP approach (dotted blue line) has an optimum at around 40-60 images, and again the DP approach (dashed green line) is relatively robust with respect to these parameters.

An overview over the results obtained using the different approaches and a comparison to the best results obtained in the ImageCLEF 2008 evaluation is given in Table 3. From the ImageCLEF evaluation, we give two results. The best-overall result from ParisTech [5] which was obtained using user interaction and manual tuning and which is therefore not directly comparable. For comparison we also give the best full-automatic result which was obtained by XRCE [1]. It can be observed that our results are slightly worse than the XRCE result with respect to P_{20} but that our results have a higher CR_{20} and a higher F-measure.

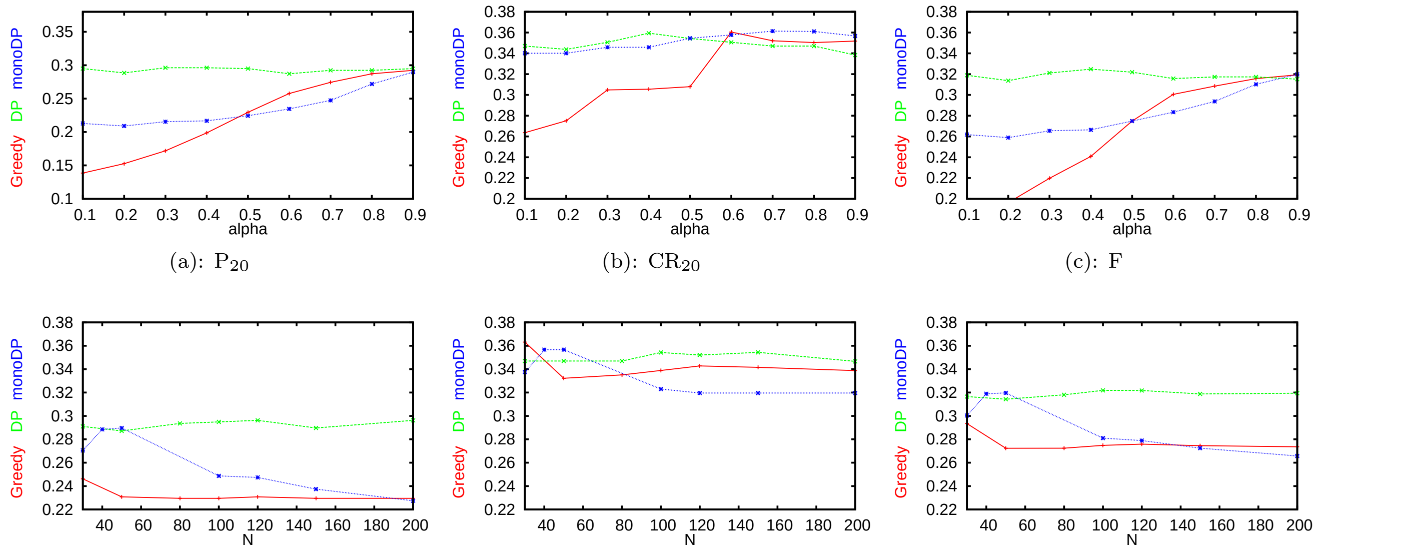
5.4 Qualitative Evaluation on the product task

In Fig. 6 we show three exemplary results on the product database. The top row of results was obtained using the conventional retrieval technique without any diversity enhancing techniques. The bottom row shows results obtained using the DP-based optimisation method described in Section 4.3. It can easily be observed that the diversity is strongly improved and the limited data in the individual concepts of this database allows to use very low α values (i.e. low weight to similarity and high weight to diversity) without finding irrelevant images.

6. DISCUSSION AND CONCLUSION

In this paper we defined a criterion to measure the diversity of results in image retrieval and proposed three approaches to directly optimising this criterion. In contrast to most other approaches to obtain diverse results in image retrieval our approaches do not build on a heuristic multi-pass architecture but optimise the result directly with respect to the defined criterion.

We have experimentally evaluated our technique quantitatively on the public ImageCLEF 2008 photo retrieval task which incorporates visual and textual information and all



(a): P_{20}

(b): CR_{20}

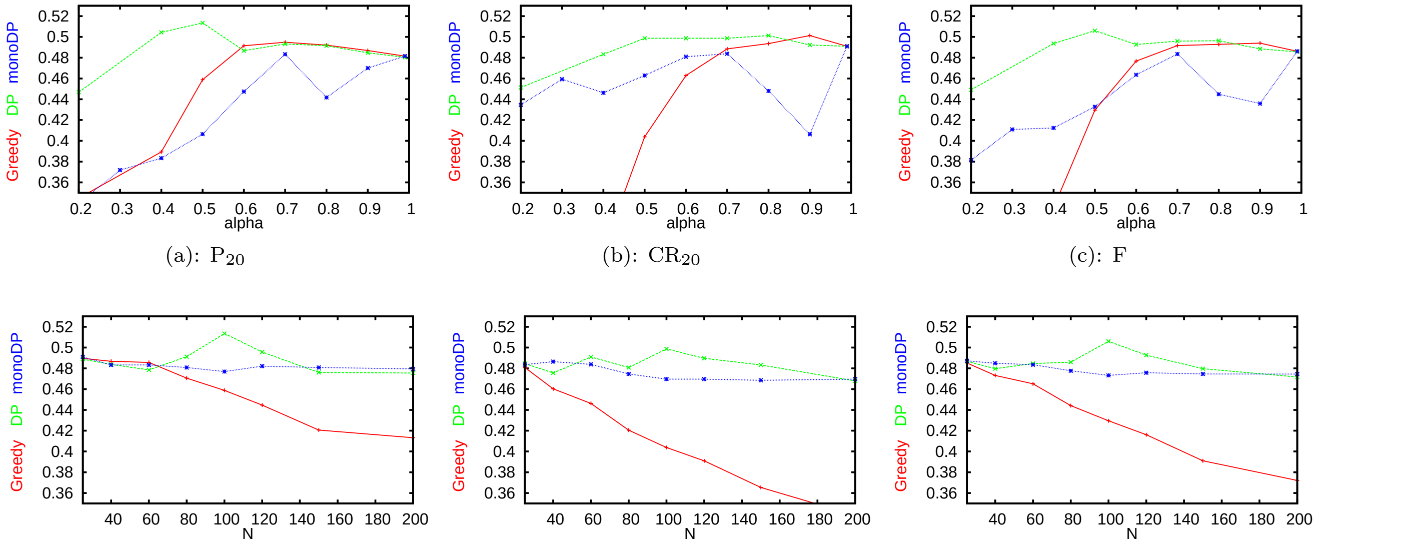
(c): F

(d): P_{20}

(e): CR_{20}

(f): F

Figure 4: The effect of different parameters on the results of the different approach in the *baseline setup*. Top row: effect of α, bottom row: effect of the number of images considered. Solid red lines are the greedy approach, dashed green lines represent the DP method and dotted blue lines are the mono-DP variant. The respective other parameter was chosen to maximise the performance.



(a): P_{20}

(b): CR_{20}

(c): F

(d): P_{20}

(e): CR_{20}

(f): F

Figure 5: The effect of different parameters on the results of the different approach in the *tuned setup* with SMART-2 text matching. Top row: effect of α, bottom row: effect of the number of images considered. Solid red lines are the greedy approach, dashed green lines represent the DP method and dotted blue lines are the mono-DP variant. The respective other parameter was chosen to maximise the performance.

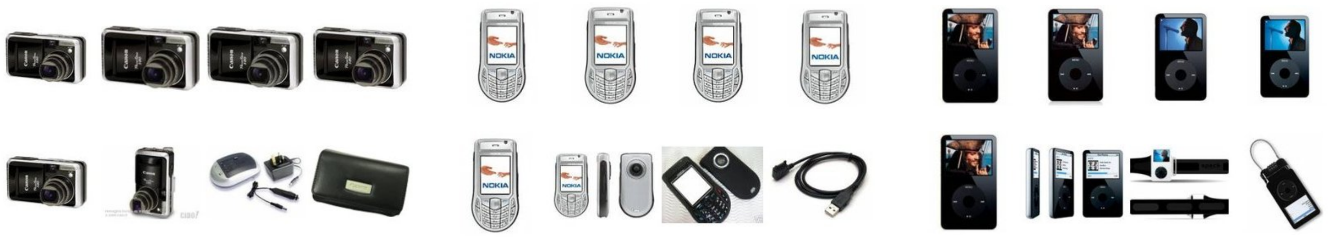


Figure 6: Three example queries on the product task. In the top row, no diversity improving method was applied and in the bottom row, the DP-based diversity-improving technique was applied.

proposed methods lead to results comparable to the state of the art or better.

Additionally we have performed a first qualitative evaluation on a novel product-search task and have seen that diverse results can be achieved in real-world tasks, too.

One interesting insight gained from our experiments is that even diverse results must be relevant otherwise the performance measures drop significantly as will user satisfaction. It was also shown that the algorithm which is based on ideas from DP algorithms is far more robust than the greedy approach and the DP-approach on the simplified criterion. In particular, it was observed that the first DP approach, which tries to optimise the full diversity criterion and therefore cannot guarantee to find the optimal solution clearly outperforms the DP approach with simplified criterion albeit the latter is guaranteed to find the optimal solution according to its criterion.

Acknowledgements. The authors would like to thank Daniel Lemire for some pointers to literature regarding diversity in his blog. The product database was provided by Exalead[2] in the framework of the QUAERO project. This work was realised as part of the QUAERO Programme, funded by OSEO, French State agency for innovation.

7. REFERENCES

[1] J. Ah-Pine, C. Cifarelli, S. Clinchant, G. Csurka, and J. Renders. XRCE's participation to ImageCLEF 2008. In *CLEF Workshop*, Aarhus, Denmark, Sept. 2008.

[2] T. Arni, P. Clough, M. Sanderson, and M. Grubinger. Overview of the ImageCLEFphoto 2008 photographic retrieval task. In *CLEF Workshop*, LNCS, Aarhus, Denmark, Sept. 2008 (printed in 2009).

[3] H. Chen and D. R. Karger. Less is more: probabilistic models for retrieving fewer relevant documents. In *SIGIR*, pp. 429–436, New York, NY, USA, 2006.

[4] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, and I. MacKinnon. Novelty and diversity in information retrieval evaluation. In *SIGIR*, pp. 659–666, New York, NY, USA, 2008.

[5] M. Ferecatu and H. Sahbi. TELECOMParisTech at ImageClefphoto 2008: Bi-modal text and image retrieval with diversity enhancement. In *CLEF Workshop*, Aarhus, Denmark, Sept. 2008.

[6] D. M. Fleder and K. Hosanagar. Recommender systems and their impact on sales diversity. In *ACM Conf. on Electronic commerce*, pp. 192–199, New York, NY, USA, 2007.

[7] M. Grubinger, P. Clough, H. Müller, and T. Deselaers. The IAPR benchmark: A new evaluation resource for visual information systems. In *Int. Conf. Language Resources and Evaluation*, Genoa, Italy, May 2006.

[8] L. Jung-Eun and A. Rong, Jin Jain. Rank-based distance metric learning: An application to image retrieval. In *CVPR*, Anchorage, AK, USA, June 2008.

[9] Q. Le and A. Smola. Direct optimization of ranking measures. online.

[10] D. Lemire, S. Downes, and S. Paquet. Diversity in open social networks. Technical report, University of Quebec, Montreal, CA, Oct. 2008.

[11] Y. Linde, A. Buzo, and R. Gray. An algorithm for vector quantization design. In *IEEE Trans. Comm.*, volume 28, pp. 84–95, Jan. 1980.

[12] L. McGinty and B. Smyth. On the role of diversity in conversational recommender systems. In *ICCBR*, 2003.

[13] S. M. McNee, J. Riedl, and J. A. Konstan. Being accurate is not enough: how accuracy metrics have hurt recommender systems. In *CHI*, pp. 1097–1101, New York, NY, USA, 2006.

[14] M. Nakazato and T. S. Huang. 3D mars: immersive virtual reality for content-based image retrieval. In *IEEE Int. Conf. on Multimedia and Expo*, pp. 44– 47, Tokyo, Japan, aug 2001.

[15] G. P. Nguyen and M. Worring. Optimization of interactive visual-similarity-based search. *ACM Trans. Multimedia Computing*, 4(1), Jan. 2008.

[16] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[17] E. Reingold, J. Nievergeld, and N. Deo. *Combinatorial algorithms: Theory and practice*. Prentice-Hall, 1977.

[18] S. Sarin and W. Kameyama. Targeting diversity in photographic retrieval task with commonsense knowledge. In *CLEF Workshop*, Aarhus, Denmark, Sept. 2008.

[19] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE TPAMI*, 22(8):888–905, Aug. 2000.

[20] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, volume 2, pp. 1470–1477, Oct. 2003.

[21] H. Tamura, S. Mori, and T. Yamawaki. Textural features corresponding to visual perception. *IEEE Trans. Systems, Man, and Cybernetics*, 8(6):460–472, June 1978.

[22] J. Tang, T. Arni, M. Sanderson, and P. Clough. Building a diversity featured search system by fusing existing tools. In *CLEF Workshop*, Aarhus, Denmark, Sept. 2008.

[23] A. Torralba, R. Fergus, and Y. Weiss. Small codes and large image databases for recognition. In *CVPR*, Anchorage, AK, USA, June 2008.

[24] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. A comparison of color features for visual concept classification. In *CIVR*, pp. 141–149, Niagara Falls, Canada, 2008.

[25] Y. Xu and H. Yin. Novelty and topicality in interactive information retrieval. *J. Am. Soc. Inf. Sci. Technol.*, 59(2):201–215, 2008.

[26] C. X. Zhai, W. W. Cohen, and J. Lafferty. Beyond independent relevance: methods and evaluation metrics for subtopic retrieval. In *SIGIR*, pp. 10–17, New York, NY, USA, 2003.

[27] M. Zhang and N. Hurley. Avoiding monotony: improving the diversity of recommendation lists. In *ACM Conf. Recommender systems*, pp. 123–130, New York, NY, USA, 2008.

[28] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *WWW*, pp. 22–32, New York, NY, USA, 2005.

[2] http://www.exalead.com/