ARTICLE IN PRESS

Information Processing and Management xxx (2012) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Dynamic two-stage image retrieval from large multimedia databases ☆

Avi Arampatzis *, Konstantinos Zagoris, Savvas A. Chatzichristofis

Department of Electrical and Computer Engineering, Democritus University of Thrace, University Campus, 67100 Xanthi, Greece

ARTICLE INFO

Article history:
Received 3 March 2011
Received in revised form 10 January 2012
Accepted 21 March 2012
Available online xxxx

Keywords:
Multimodal retrieval
Multimedia retrieval
Image retrieval
Fusion

ABSTRACT

Content-based image retrieval (CBIR) with global features is notoriously noisy, especially for image queries with low percentages of relevant images in a collection. Moreover, CBIR typically ranks the whole collection, which is inefficient for large databases. We experiment with a method for image retrieval from multimedia databases, which improves both the effectiveness and efficiency of traditional CBIR by exploring secondary media. We perform retrieval in a two-stage fashion: first rank by a secondary medium, and then perform CBIR only on the top-*K* items. Thus, effectiveness is improved by performing CBIR on a 'better' subset. Using a relatively 'cheap' first stage, efficiency is also improved via the fewer CBIR operations performed. Our main novelty is that *K* is dynamic, i.e. estimated per query to optimize a predefined effectiveness measure. We show that our dynamic two-stage method can be significantly more effective and robust than similar setups with static thresholds previously proposed. In additional experiments using local feature derivatives in the visual stage instead of global, such as the emerging visual codebook approach, we find that two-stage does not work very well. We attribute the weaker performance of the visual codebook to the enhanced visual diversity produced by the textual stage which diminishes codebook's advantage over global features. Furthermore, we compare dynamic two-stage retrieval to traditional score-based fusion of results retrieved visually and textually. We find that fusion is also significantly more effective than single-medium baselines. Although, there is no clear winner between two-stage and fusion, the methods exhibit different robustness features; nevertheless, two-stage retrieval provides efficiency benefits over fusion.

1. Introduction

In content-based image retrieval (CBIR), images are represented by global or local features. Global features are capable of generalizing an entire image with a single vector, describing color, texture, or shape. Local features are computed at multiple points on an image and are capable of recognizing objects.

CBIR with global features is notoriously noisy for image queries of low *generality*, i.e. the fraction of relevant images in a collection. In contrast to text retrieval where documents matching no query keyword are not retrieved, CBIR methods typically rank the whole collection via some distance measure. For example, a query image of a red tomato on white background would retrieve a red pie-chart on white paper. If the query image happens to have a low generality, early rank positions may be dominated by spurious results such as the pie-chart, which may even be ranked before tomato images on non-white backgrounds. Fig. 3a and b demonstrate this particular problem.

☆ A preliminary version of parts of this work was presented in ECIR (European Conference on Information Retrieval) 2011.

* Corresponding author. Tel./fax: +30 25410 79513.
E-mail addresses: avi@ee.duth.gr (A. Arampatzis), kzagoris@ee.duth.gr (K. Zagoris), schatzic@ee.duth.gr (S.A. Chatzichristofis).

0306-4573/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ipm.2012.03.005

Local-feature approaches provide a slightly better retrieval effectiveness than global features (Aly, Welinder, Munich, & Perona, 2009). They represent images with multiple points in a feature space in contrast to single-point global feature representations. While local approaches provide more robust information, they are more expensive computationally due to the high dimensionality of their feature spaces and usually need nearest neighbors approximation to perform points-matching (Popescu, Moëllic, Kanellos, & Landais, 2009). High-dimensional indexing still remains a challenging problem in the database field. Thus, global features are more popular in CBIR systems as they are easier to handle and still provide basic retrieval mechanisms. In any case, CBIR with either local or global features does not scale up well to large databases efficiency-wise. In small databases, a simple sequential scan may be acceptable, however, scaling up to millions or billion images efficient indexing algorithms are imperative (Li, Chen, Zhang, Lin, & Ma, 2006).

Nowadays, information collections are not only large, but they also consist of *multimedia*. Take as an example Wikipedia, where each topic or article consists of a textual part and may include non-textual media such as image, sound, and video. Furthermore, such collections may also provide several ways of accessing each medium. For example, a topic may be covered in several natural languages, and non-textual media may be annotated in a variety of metadata fields, such as filename, author, description, and comment, possibly also in several languages. Without a clear definition and rather confusingly, some have come to use the term *multimodal* interchangeably to multimedia for describing such retrieval setups.

Conforming to one of the definitions of modality in the dictionary, i.e. *"a quality, attribute, or circumstance that denotes mode, mood, or manner"*,[1] we find it more appropriate in IR to define modality as a manner of retrieval. For example, a multimedia collection of two media, e.g. text and image, may consist of more than two modalities, e.g. English text, French text, image texture, image color, etc., for each of its items. Thus, in the rest of this paper, we will talk of multimodal retrieval as a more general (but at the same time more granular) term which may also imply—but not necessarily denote—multimedia. Current search engines usually focus on limited numbers of modalities, e.g. English text queries on English articles or maybe on annotations of other media as well, not making use of all the available information. In an image retrieval system where users are assumed to target visual similarity, all media beyond image can be considered as secondary; nevertheless, modalities from secondary media can still provide useful information for improving image retrieval.

In this paper, we experiment with a method for image retrieval from large multimedia databases, which targets to improve both the effectiveness and efficiency of traditional CBIR by exploring information from secondary media. In the setup considered, an information need is expressed by a query in the primary medium (i.e. an image example) accompanied by a query in a secondary medium (e.g. text). The core idea for improving effectiveness is to raise query generality before performing CBIR, by reducing collection size via filtering methods. In this respect, we perform retrieval in a two-stage fashion: first use the secondary medium to rank the collection and then perform CBIR only on the top-K items. Using a 'cheaper' secondary medium, this improves also efficiency by cutting down on costly CBIR operations.

Best results re-ranking by visual content has been seen before, but mostly in different setups than the one we consider or for different purposes, e.g. result clustering (Barthel, 2008) or diversity (van Leuken, Pueyo, Olivares, & van Zwol, 2009). Others used external information, e.g. an external set of diversified images (Popescu et al., 2009) (also, they did not use image queries), web images to depict a topic (Myoupo, Popescu, Le Borgne, & Moellic, 2009), or training data (Berber & Alpkocak, 2009). All these approaches, as well as (Maillot, Chevallet, & Lim, 2006), employed a static predefined K for all queries, except (Popescu et al., 2009) who re-ranked the top-30% of retrieved items. Most of them used global features for images. Effectiveness results have been mixed; it worked for some, it did not for others, while some did not provide a comparative evaluation or system-study. Later, we will review the related literature in more detail.

In view of the related literature, our main contributions are the following. First, our threshold is calculated *dynamically* per query to optimize a predefined effectiveness measure, without using external information or training data. Second, we provide an extensive evaluation in relation to thresholding types, levels, and robustness. Third, we investigate the influence of different effectiveness levels of the second visual stage on the whole two-stage procedure. Fourth, beyond using global features, we also investigate the performance of the proposed setup for local feature derivatives in the visual stage. Fifth, we provide a comprehensive review of related literature and discuss the conditions under which such setups can be applied effectively. Traditionally, the method that has been followed in order to deal effectively with multimodal databases is to search the modalities separately and fuse their results, e.g. with a linear combination of the retrieval scores of all modalities per item; the same can be done across media. Consequently, our sixth contribution is a theoretical and experimental comparison of two-stage to fusion.

The rest of the paper is organized as follows. In Section 2 we discuss the assumptions, hypotheses, and requirements behind two-stage image retrieval from multimedia databases. In Section 3 we perform an experiment on a standardized multimodal snapshot of Wikipedia. In Section 4 we switch the type of features used for the visual stage from global to locally-based and repeat a part of the main experiment. In Section 5 we compare the proposed two-stage setup to fusion. In Section 6 we review related work. Conclusions and directions for further research are summarized in Section 7.

[1] http://www.thefreedictionary.com/modality (accessed on 18.02.11).

Please cite this article in press as: Arampatzis, A., et al. Dynamic two-stage image retrieval from large multimedia databases. *Information Processing and Management* (2012), http://dx.doi.org/10.1016/j.ipm.2012.03.005

2. Two-stage image retrieval from multimedia databases

Mutlimedia databases consist of multiple media for each retrievable item; in the setup we consider these are image and annotations. On the one hand, the visual content of images corresponds to large amounts of information which can hardly be described by words. On the other hand, textual descriptions are key to retrieving relevant results for a query but at the same time capture little visual information (van Leuken et al., 2009); past experiments have shown that users tend to annotate images based on the objects appearing in them rather than color or texture (Mulhem & Lim, 2002).

Traditionally, the method that has been followed in order to deal effectively with multimodal databases is to search the modalities separately and fuse their results, e.g. with a linear combination of the retrieval scores of all modalities per item. While fusion has been proved robust, we argue that it has a couple of important issues:

- Appropriate weighing of modalities and score normalization/combination are not a trivial problems and may require training data.
- It is not a theoretically sound method if results are assessed by visual similarity only; the influence of textual scores may worsen the visual quality of end-results.

The latter issue points to that there is a *primary medium*, i.e. the one targeted and assessed by users. Additionally, the total search time in fusion is the sum of the times taken for searching the participating modalities.

An approach that may tackle the issues of fusion would be to search in a two-stage fashion: first rank with a secondary medium, draw a rank-threshold, and then re-rank only the top items with the primary medium. The assumption on which such a two-stage setup is based on is the existence of a primary medium, and the success would largely depend on the *relative retrieval effectiveness* of the two media involved. For example, if text retrieval always performs better than CBIR (irrespective of query generality), then CBIR is redundant. If it is the other way around, only CBIR will be sufficient. Thus, the hypothesis is that CBIR can do better than text retrieval in small sets or sets of high query generality.

In order to reduce collection size raising query generality, a ranking can be thresholded at an arbitrary rank or item score. This improves the efficiency by cutting down on costly CBIR operations, but it may not improve too much the result quality: a too tight threshold would produce similar results to a text-only search making CBIR redundant, while a too loose threshold would produce results haunted by the red-tomato/red-pie-chart effect mentioned in the Introduction. Three factors determine what the right threshold is:

1. the number of relevant items in the collection,
2. the quality of the ranking, and
3. the measure that the threshold targets to optimize (Robertson & Hull, 2000).

The first two factors are query-dependent, thus thresholds should be selected *dynamically* per query, not statically as most previously proposed methods in the literature (reviewed in Section 6).

The approach of Popescu et al. (2009), who re-rank the top-30% retrieved items which can be considered dynamic, does not take into account the three aforementioned factors. While the number of retrieved results might be argued correlated to the number of relevant items (thus, seemingly taking into account the first factor), this correlation can be very weak at times, e.g. consider a high frequency query word (almost a stop-word) which would retrieve large parts of the collection. Further, such percentage thresholding seems remotely-connected to factors (2) and (3). Consequently, we will resort to the approach of Arampatzis, Kamps, and Robertson (2009) which, based on the distribution of item scores, is capable of estimating (1), as well as mapping scores to probabilities of relevance. Having the latter, (2) can be determined, and any measure defined in (3) can be optimized in a straightforward way. More on the method can be found in the last-cited study.

Targeting to enhance query generality, the most appropriate measure to optimize would be precision. However, since the *smoothed* precision estimated by the method of Arampatzis et al. (2009) monotonically declines with rank, it makes sense to set a precision threshold. The choice of precision threshold is dependent on the effectiveness of the CBIR stage: it can be seen as guaranteeing the minimum generality required by the CBIR method at hand for achieving good effectiveness. Not knowing the relation between CBIR effectiveness and minimum required generality, we will try a series of thresholds on precision, as well as, to optimize other cost-gain measures. Thus, while it may seem that we exchange the initial problem of where to set a static threshold with where to threshold precision or which measure to optimize, it will turn out that the latter problem is less sensitive to its available options, as we will see.

A possible drawback of the two-stage setup considered is that relevant images with empty or very noisy secondary media would be completely missed, since they will not be retrieved by the first stage. If there are any improvements compared to single-stage text-only or image-only setups, these will first show up on early precision since only the top results are re-ranked; mean average precision or other measures may improve as a side effect. Fusion does not have these problems. In any case, there are efficiency benefits from searching the most expensive medium only on a subset of the collection.

The requirement of such a two-stage CBIR at the user-side is that information needs are expressed by visual as well as textual descriptions. The community is already experimenting with such setups, e.g. the ImageCLEF 2010 Wikipedia Retrieval task was performed on a multimodal collection with topics made of textual and image queries at the same time (Popescu

Please cite this article in press as: Arampatzis, A., et al. Dynamic two-stage image retrieval from large multimedia databases. *Information Processing and Management* (2012), http://dx.doi.org/10.1016/j.ipm.2012.03.005

et al., 2010). Past experiments have shown that users seem to find it more intuitive to access images using natural language descriptions rather than low-level characteristics such as colors or textures (Martinet, Chiaramella, & Mulhem, in press; Rodden & Wood, 2003). Furthermore, multimodal or holistic query interfaces are showing up in experimental search engines allowing concurrent multimedia queries (Zagoris et al., 2010). As a last resort, automatic image annotation methods (Chang, Goh, Sychay, & Wu, 2003; Li & Wang, 2008) may be employed for generating queries for secondary media in traditional image retrieval systems.

3. Experiments on Wikipedia

In this section, we report on experiments performed on a standardized multimodal snapshot of Wikipedia. It is worth noting that the collection is one of the largest benchmark image databases for today's standards. It is also highly heterogeneous, containing color natural images, graphics, grayscale images, etc., in a variety of sizes.

3.1. Datasets, systems, and methods

The ImageCLEF 2010 Wikipedia test collection has image as its primary medium, consisting of 237,434 items, associated with noisy and incomplete user-supplied textual annotations and the Wikipedia articles containing the images. Associated annotations exist in any combination of English, German, French, or any other unidentified (non-marked) language. There are 70 test topics, each one consisting of a textual and a visual part: three title fields (one per language—English, German, French), and one or more example images. The topics are assessed by visual similarity to the image examples. More details on the dataset can be found in Popescu et al. (2010).

For text indexing and retrieval, we employ the Lemur Toolkit V4.11 and Indri V2.11 with the tf.idf retrieval model.[2] We use the default settings that come with these versions of the system except that we enable Krovetz stemming. We index only the English annotations, and use only the English query of the topics.

Compact Composite Descriptors (CCDs) (Chatzichristofis, Arampatzis, & Boutalis, 2010; Chatzichristofis, Zagoris, Boutalis, & Papamarkos, 2010) are global image features, capturing more than one type of information at the same time in a very compact representation. We index the images with two CCDs: the Joint Composite Descriptor (JCD) and the Spatial Color Distribution (SpCD). The JCD is developed for color natural images and combines color and texture information (Chatzichristofis, Arampatzis, et al., 2010). In several benchmarking databases, JCD has been found more effective than MPEG-7 descriptors (Chatzichristofis, Arampatzis, et al., 2010). The SpCD combines color and its spatial distribution; it is considered more suitable for colored graphics since they consist of a relatively small number of colors and less texture regions than color natural images. It is recently introduced in Chatzichristofis, Boutalis, and Lux (2010) and found to perform better than JCD in a heterogeneous image database (Chatzichristofis & Arampatzis, 2010).

We evaluate on the top-1000 results with mean average precision (MAP), precision at 10 and 20, and bpref (Buckley & Voorhees, 2004). The bpref measure is inversely related to the fraction of judged non-relevant documents that are retrieved before relevant documents (Buckley & Voorhees, 2004); adding additional unjudged documents to a retrieved set can have no effect on that sets bpref score, but can have significant influence on the other measures scores. We are using it as a measure of ground-truth incompleteness that signals whether we retrieve un-judged items.

3.2. Thresholding and re-ranking

We investigate two types of thresholding: static and dynamic. In static thresholding, the same fixed pre-selected rank threshold K is applied to all topics. We experiment with levels of K at 25, 50, 100, 250, 500, and 1000. The results that are not re-ranked by image are retained as they are ranked by text, also in dynamic thresholding.

For dynamic thresholding, we use the Score-Distributional Threshold Optimization (SDTO) as described in Arampatzis et al. (2009). For tf.idf scores, we used the *technically truncated* model of a normal-exponential mixture. The method normalizes retrieval scores to probabilities of relevance (prels), enabling the optimization of K for any user-defined effectiveness measure. Per query, we search for the optimal K in $[0, 2500]$, where 0 or 1 results to no re-ranking. Thus, for estimation with the SDTO we truncate at the score corresponding to rank 2500 but use no truncation at high scores as tf.idf has no theoretical maximum. If there are 25 text results or less, we always re-rank by image; these are too few scores to apply the SDTO reliably. In this category fall the topics 1, 10, 23, and 46, with only 18, 16, 2, and 18 text results respectively. The biggest strength of the SDTO is that it does not require training data; more details on the method can be found in the last-mentioned study.

We experiment with the SDTO by thresholding on prel as well as on precision. Thresholding on fixed prels happens to optimize *linear utility measures* (Lewis, 1995), with corresponding rank thresholds:

- $\max K$: $P(\text{rel}|D_K) > \theta$, where D_K is the Kth ranked document. For the prel threshold θ, we try six values. Two of them are:
  - $\theta = 0.5000$: It corresponds to 1 loss per relevant non-retrieved and 1 loss per non-relevant retrieved, i.e. the Error Rate, and it is precision-recall balanced.

[2] http://www.lemurproject.org.

Please cite this article in press as: Arampatzis, A., et al. Dynamic two-stage image retrieval from large multimedia databases. *Information Processing and Management* (2012), http://dx.doi.org/10.1016/j.ipm.2012.03.005

– $\theta = 0.3333$: It corresponds to 2 gain per relevant retrieved and 1 loss per non-relevant retrieved, i.e. the T9U measure used in the TREC 2000 Filtering Track (Robertson & Hull, 2000), and it is recall-oriented.

These prel thresholds may optimize other measures as well; for example, 0.5000 optimizes also the utility measure of 1 gain per relevant retrieved and 1 loss per non-relevant retrieved. Thus, irrespective of which measure prel thresholds optimize, we arbitrarily enrich the experimental set of levels with four more thresholds: 0.9900, 0.9500, 0.8000, and 0.1000.

Furthermore, having normalized scores to prels, we can estimate precision in any top-K set by simply adding the prels and dividing by K. The estimated precision can be seen as the generality in the sub-ranking. According to the hypothesis that the effectiveness of CBIR is positively correlated to query generality, we experiment with the following thresholding:

- max$K$: Prec@$K > g$, where for g is the minimum generality required by the CBIR at hand for good effectiveness. Having no clue on usable g values, we arbitrarily try levels of g at 0.9900, 0.9500, 0.8000, 0.5000, 0.3333, and 0.1000.

3.3. Fusion of image modalities

In the current setup, we index images with more than one descriptor. Moreover, most topics have more than one example query image. Thus, we briefly describe here an appropriate method for fusing the image modalities resulting from the multiple descriptors and example images.

Let i be the index running over example images ($i = 1,2,\ldots$) and j running over the visual descriptors ($j \in \{1,2\}$). Thus, DESC_{ji} is the score of a collection item against the ith example image for the jth descriptor. We normalize DESC_{ji} values with MinMax, taking the maximum score seen across example images per descriptor. Assuming that the descriptors capture orthogonal information, we add their scores per example image. Then, to take into account all example images, the natural combination is to assign to each collection image the maximum similarity seen from its comparisons to all example images; this can be interpreted as looking for images similar to *any* of the example images. Summarizing, the score s for a collection image against the topic is defined as:

$$s = \max_i \left(\sum_j \mathrm{MinMax}(\mathrm{DESC}_{ji}) \right) \tag{1}$$

3.4. Setting the baseline

In initial experiments, we investigated the effectiveness of each of the stages individually, trying to tune them for best results.

In the textual stage, we employ the tf.idf model since it has been found to work well with the SDTO (Arampatzis, Robertson, & Kamps, 2009). The SDTO method fits a binary mixture of probability distributions on the score distribution (SD). A previous study suggested that while long queries tend to lead to smoother SDs and improved fits, threshold predictions are better for short queries of high quality keywords (Arampatzis et al., 2009). To be on the safe side, in initial experiments we tried to increase query length by enabling pseudo relevance feedback of the top-10 documents, but all our combinations of the parameter values for the number of feedback terms and initial query weight led to significant decreases in the effectiveness of text retrieval. We attribute this to the noisy nature of the annotations. Consequently, we do not run any two-stage experiments with pseudo relevance feedback at the first textual stage.

In the visual stage, first we tried the JCD alone, as the collection seems to contain more color natural images than graphics, and used only the first example image; this represents a simple but practically realistic setup. Then, incorporating all example images but still using only the JCD, we used Eq. (1) which in this case simplifies to $s = \max_i \mathrm{JCD}_i$. Last, employing all available example images and descriptors, we used Eq. (1) as given above. Table 1 presents the results.

The image-only runs perform far below the text-only run. This puts in perspective the quality of the currently effective global CBIR descriptors: their effectiveness in image retrieval is much worse than the effectiveness of the traditional tf.idf text retrieval model even on sparse and noisy annotations. Since the image-only runs would have provided very weak baselines, we choose as a much stronger baseline for statistical significance testing the text-only run. This makes sense also from an efficiency point of view: if using a secondary text modality for image retrieval is more effective than current CBIR methods, then there is no reason at all for using computationally costly CBIR methods.

Table 1
Effectiveness of different CBIR setups against tf.idf text-only retrieval.

Item scoring by	MAP	P@10	P@20	bpref
JCD_1	.0058	.0486	.0479	.0352
$\max_i \mathrm{JCD}_i$	.0072	.0614	.0614	.0387
Eq. (1)	.0107	.0871	.0871	.0402
tf.idf (text-only)	.1293	.3614	.3314	.1806

Please cite this article in press as: Arampatzis, A., et al. Dynamic two-stage image retrieval from large multimedia databases. *Information Processing and Management* (2012), http://dx.doi.org/10.1016/j.ipm.2012.03.005

Table 2
Two-stage image retrieval results. The best results per measure and thresholding type are in boldface. Significance-tested with a bootstrap test, one-tailed, at significance levels 0.05 ($^{\triangle\triangledown}$), 0.01 ($^{\blacktriangle\triangledown}$), and 0.001 ($^{\blacktriangle\blacktriangledown}$), against the text-only baseline.

threshold		$\widetilde{K}$	JCD$_1$ MAP	JCD$_1$ P@10	JCD$_1$ P@20	JCD$_1$ bpref	Equation 1 MAP	Equation 1 P@10	Equation 1 P@20	Equation 1 bpref
text-only		—	.1293	.3614	.3307	.1809	.1293	.3614	.3307	.1809
K	25	25	**.1162**$^{\triangledown}$	**.3957**	.3457$^{\triangle}$	.1641$^{\triangledown}$	**.1170**	.3957	.3464$^{\triangledown}$	.1661$^{\triangledown}$
	50	50	.1144$^{\triangledown}$	.3829	**.3579**$^{\triangle}$	.1608$^{\triangledown}$	.1155	**.4000**	.3543	.1649$^{\triangledown}$
	100	100	.1138	.3786	.3471	.1609$^{\triangledown}$	.1133	.3914	.3471	.1626$^{\triangledown}$
	250	250	.1081$^{\triangledown}$	.3414	.3164	**.1644**	.1090$^{\triangledown}$	.3786	**.3550**	**.1662**
	500	500	.0968$^{\triangledown}$	.3200	.3007	.1575$^{\triangledown}$	.0993$^{\triangledown}$	.3500	.3200	.1588$^{\triangledown}$
	1000	1000	.0865$^{\blacktriangledown}$	.2871$^{\triangledown}$	.2729$^{\triangledown}$	.1493$^{\triangledown}$	.0904$^{\blacktriangledown}$	.3300	.3014	.1507$^{\triangledown}$
g	.9900	49	**.1364**	**.4214**$^{\triangle}$	.3550	.1902$^{\triangle}$	.1287	.4286$^{\triangle}$	.3721$^{\triangle}$	.1831
	.9500	68	.1352	.4171$^{\triangle}$	**.3586**	**.1912**$^{\triangle}$	**.1307**	**.4343**$^{\triangle}$	**.3786**$^{\triangle}$	**.1850**
	.8000	95	.1318	.4000	.3536	.1892	.1258	.4257$^{\triangle}$	.3693	.1815
	.5000	151	.1196	.3814	.3393	.1808	.1120$^{\triangledown}$	.3986	.3350	.1701
	.3333	237	.1085$^{\triangledown}$	.3500	.3000	.1707	.1029$^{\triangledown}$	.3657	.3157	.1619$^{\triangledown}$
	.1000	711	.0864$^{\blacktriangledown}$	.2871$^{\triangledown}$	.2621$^{\triangledown}$	.1461$^{\blacktriangledown}$	.0836$^{\blacktriangledown}$	.3143	.2779$^{\triangledown}$	.1428$^{\blacktriangledown}$
θ	.9900	42	.1342	.4043	.3414	.1865	.1376$^{\triangle}$	.4286$^{\triangle}$	.3714$^{\triangle}$	.1899$^{\triangle}$
	.9500	51	.1371	.4214$^{\triangle}$	.3586	.1903$^{\triangle}$	.1390$^{\triangle}$	.4314$^{\triangle}$	.3771$^{\triangle}$	.1917$^{\triangle}$
	.8000	81	**.1384**$^{\triangle}$	**.4229**$^{\triangle}$	.3614	.1921$^{\triangle}$	**.1428**$^{\triangle}$	**.4443**$^{\triangle}$	**.3857**$^{\triangle}$	**.1959**$^{\triangle}$
	.5000	91	.1367	.4057	.3571	.1919$^{\triangle}$	.1405$^{\triangle}$	.4357$^{\triangle}$	.3821$^{\triangle}$	.1943$^{\triangle}$
	.3333	109	.1375	.4129	**.3636**	**.1933**$^{\triangle}$	.1403$^{\triangle}$	.4357$^{\triangle}$	.3807$^{\triangle}$	.1942$^{\triangle}$
	.1000	130	.1314	.4100	.3629	.1866	.1373	.4329$^{\triangle}$	.3850$^{\triangle}$	.1921
image-only		—	.0058$^{\blacktriangledown}$	.0486$^{\blacktriangledown}$	.0479$^{\blacktriangledown}$	.0352$^{\blacktriangledown}$	.0107$^{\blacktriangledown}$	.0871$^{\blacktriangledown}$	.0871$^{\blacktriangledown}$	.0402$^{\blacktriangledown}$

Comparing the image-only runs to each other, we see that using more information—either from more example images or more descriptors—improves effectiveness. In order to investigate the impact of the effectiveness level of the second stage on the whole two-stage procedure, we will present two-stage results for both the best and the worst CBIR methods.

3.5. Experimental results

Table 2 presents two-stage image retrieval results against text- and image-only retrieval. It is easy to see that the dynamic thresholding methods improve retrieval effectiveness in most of the experiments. Especially, dynamical thresholding using θ shows improvements for all values we tried. The greatest improvement (+23%) is observed in P@10 for $\theta = 0.8$. The table con-

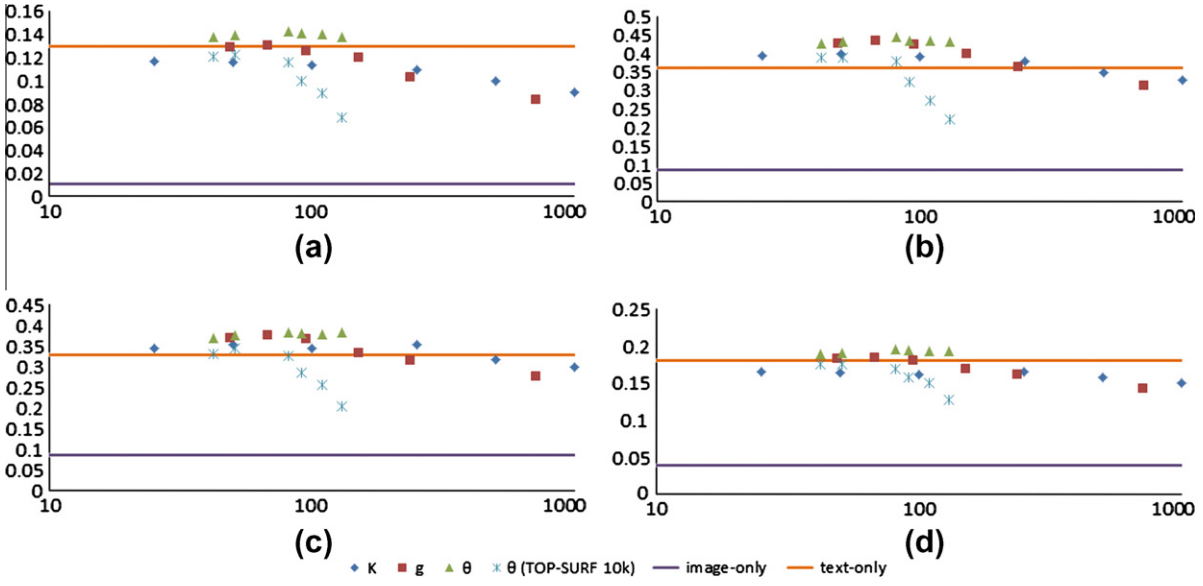

Fig. 1. Effectiveness, for the strongest CBIR stage: (A) MAP, (B) P@10, (C) P@20, (D) bpref.

Please cite this article in press as: Arampatzis, A., et al. Dynamic two-stage image retrieval from large multimedia databases. *Information Processing and Management* (2012), http://dx.doi.org/10.1016/j.ipm.2012.03.005

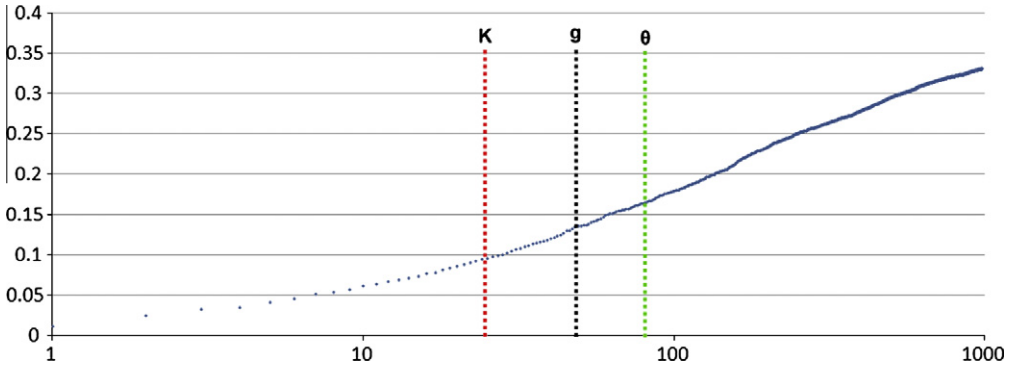

Fig. 2. Macro-averaged recall against rank.

tains lots of numbers; while there may be consistent increases or decreases in some places, in the rest of this section we focus and summarize only the statistically significant differences.

Irrespective of measure and CBIR method, the best thresholds are roughly at: 25 or 50 for K, 0.95 for g, and 0.8 for θ. The weakest thresholding method is the static K: there are very few improvements only in P@20 at tight cutoffs, but they are accompanied by a reduced MAP and bpref. Actually, static thresholds hurt MAP and/or bpref almost anywhere. Effectiveness degrades also in early precision for $K = 1000$. Dynamic thresholding is much more robust. Comparing the two CBIR methods at the second stage, the stronger method helps the dynamic methods considerably while static thresholding does not seem to receive much improvement.

Concerning the dynamic thresholding methods, the probability thresholds θ correspond to tighter *effective* rank thresholds than these of the precision thresholds g, for g and θ taking values in the range [0.1000, 0.9900]. As a proxy for the effective K we use the median threshold $\widetilde{K}$ across all topics. This is expected since precision declines slower than prel. Nevertheless, the fact that a wide range of prel thresholds results to a tight range of $\widetilde{K}$, reveals a sharp decline in prel below some score per query. This makes the end-effectiveness less sensitive to prel thresholds in comparison to precision thresholds, thus more robust against possibly unsuitable user-selected values. Furthermore, if we compare the dynamic methods at similar $\widetilde{K}$, e.g. $g = 0.9900$ to $\theta = 0.9500$ ($\widetilde{K} \approx 50$) and $g = 0.8000$ to $\theta = 0.5000$ ($\widetilde{K} \approx 93$), we see that prel thresholds perform slightly better. Fig. 1 depicts the evaluation measures against $\widetilde{K}$ for all methods and the stronger CBIR; Fig. 3b–e present the top image results for the query of Fig. 3a. We will explain in Section 4 what "TOP-SURF 10k" and Fig. 3f are.

As we mentioned in Section 2, a possible drawback of the two-stage setup is that relevant images with empty or very noisy textual descriptions would be completely missed, since they will not be retrieved by the first stage, possibly hurting recall. Fig. 2 shows the macro-averaged recall against rank, as well as the best K, g and θ thresholds. It can be seen that recall hardly reaches 0.35 at top-1000; since the average relevant items per topic is 252.3 we can conclude that the text-only stage is far from perfect and that the aforementioned drawback has some merit.

In summary, static thresholding improves initial precision at the cost of MAP and bpref, while dynamic thresholding on precision or prel does not have this drawback. The choice of a static or precision threshold influences greatly the effectiveness, and unsuitable choices (e.g. too loose) may lead to a degraded performance. Prel thresholds are much more robust in this respect. As expected, better CBIR at the second stage leads to overall improvements, nevertheless, the thresholding type seems more important: While the two CBIR methods we employ vary greatly in performance (the best has almost double the effectiveness of the other), static thresholding is not influenced much by this choice; we attribute this to its lack of respect for the number of relevant items and for the ranking quality. Dynamic methods benefit more from improved CBIR. Overall, prel thresholds perform best, for a wide range of values.

4. Two-stage with bag-of-visual-words

So far, the hypothesis for effective two-stage image retrieval was that CBIR can do better than text retrieval in small or high query generality collections; a hypothesis deemed valid by the experiments in Section 3, at least for global features such as the CCDs. In this section, we investigate whether the hypothesis holds also for other types of features beyond global.

SURF local features are among the best-interest-point descriptors currently available. They have been shown to outperform other well-known methods based on interest points, such as SIFT (Lowe, 2004) and GLOH (Mikolajczyk & Schmid, 2005). Nevertheless, in large-scale CBIR, it is clear that using the SURF descriptors is storage-wise infeasible (Thomee, Bakker, & Lew, 2010).

Bag-of-visual-words (BOVW) (Cula & Dana, 2001) is a representation of images which is built using a large set of local features. They are inspired by the bag-of-words models in text classification, where a document is represented by a set of distinct keywords. Analogously, in BOVW models, an image is represented by a set of distinct visual words derived from local features. BOVWs are fast becoming a widely used representation for content-based image retrieval, for mainly two reasons:

Please cite this article in press as: Arampatzis, A., et al. Dynamic two-stage image retrieval from large multimedia databases. *Information Processing and Management* (2012), http://dx.doi.org/10.1016/j.ipm.2012.03.005

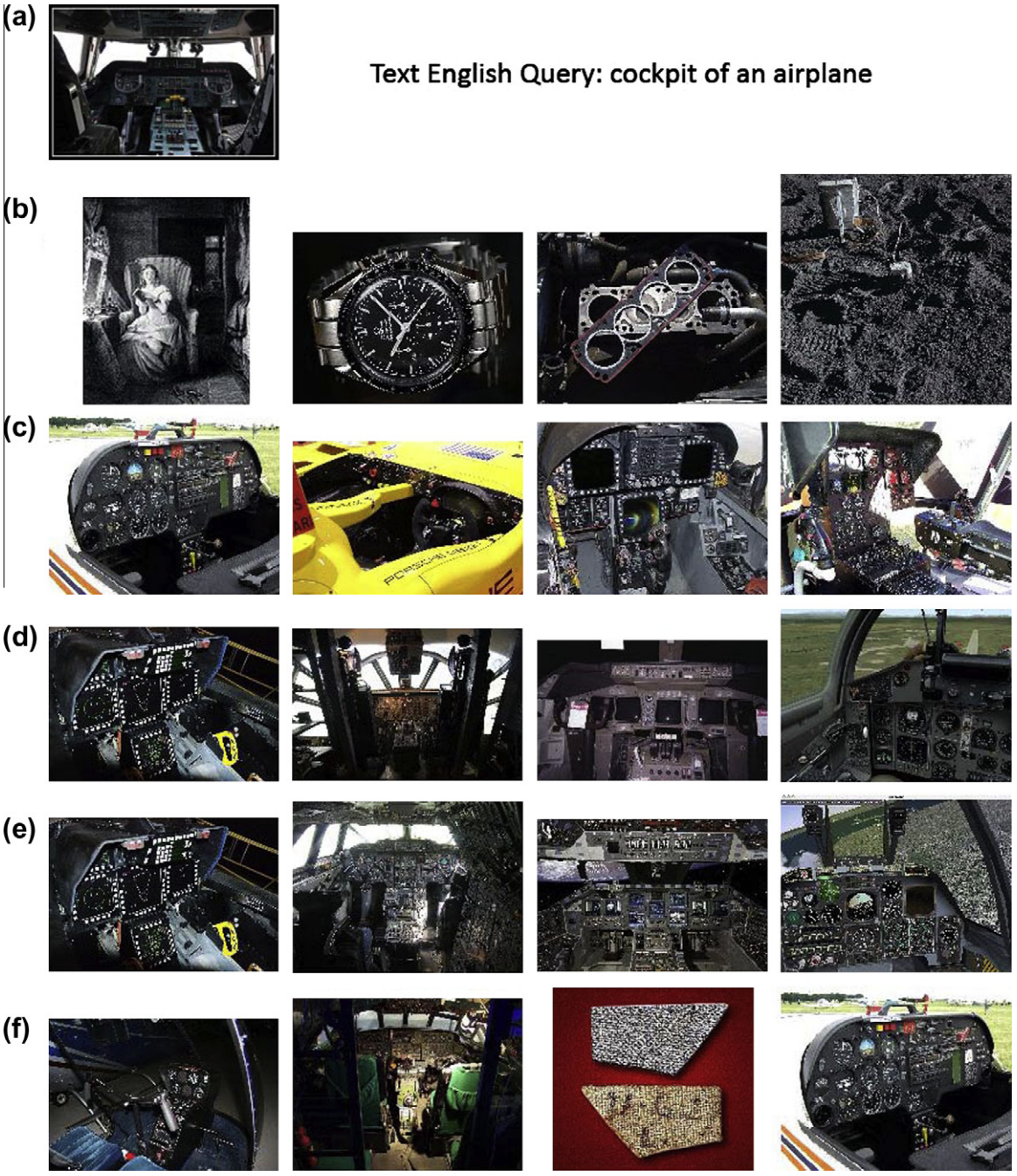

Fig. 3. Retrieval results: (a) query, (b) image-only, (c) text-only, (d) $K = 25$, (e) $\theta = 0.8$, (f) $\theta = 0.8$ (TOP-SURF 10 k).

their better retrieval effectiveness over global feature representations on near identical images, and much better efficiency than local feature representations. However, experimental results of reported work show that the commonly generated visual words are still not as expressive as the text words (Zhang, Tian, Hua, Huang, & Li, 2009).

The most modern implementation of BOVW suitable for a wide range of CBIR applications is the TOP-SURF (Thomee et al., 2010) descriptor. TOP-SURF combines interest points with visual words, resulting in a high performance compact descriptor. The TOP-SURF descriptor, initially extracts SURF local features from the images and then groups these features into a desired number of clusters. Each cluster can be seen as a visual word. All visual words are stored in a visual dictionary. Next, the tf.idf weighting is applied in order to assign a score to all the visual words in the histogram. The TOP-SURF image descriptor is created by choosing the top scoring visual words.

Please cite this article in press as: Arampatzis, A., et al. Dynamic two-stage image retrieval from large multimedia databases. *Information Processing and Management* (2012), http://dx.doi.org/10.1016/j.ipm.2012.03.005

4.1. TOP-SURF vs. CCDs

We index the images with the TOP-SURF descriptor, employing two visual-word dictionaries: one with 10,000 and the other with 200,000 visual words. In order to investigate the impact of dictionary size on both the retrieval effectiveness and the matching time. For CBIR, we used the JCD, SpCD, and TOP-SURF, separately, as well as the late fusion setup of JCD and SpCD as in Eq. (1). The general setup is similar to the one described in Section 3. For measuring efficiency, we report the average matching time per topic. The results are presented in Table 3.

For all θ, the CCDs perform similarly (JCD) or significantly better (SpCD and JCD/SpCD) than the text-only baseline, while the TOP-SURF descriptor shows significant drops in effectiveness irrespective of dictionary size. The differences in effectiveness of CCDs and TOP-SURF are larger in early precision than in MAP. We also observe that the TOP-SURF effectiveness degrades with increased dictionary size. Furthermore, TOP-SURF are more sensitive to the choice of θ: as θ decreases (i.e. for larger Ks), effectiveness deteriorates faster than this of the CCDs. Efficiency-wise, the experimental results show that although the TOP-SURF uses a speedy matching algorithm, it still cannot match the speed of the CCDs.

A possible explanation for the inferior effectiveness of BOVW in the proposed setup is the following. Although BOVW models have the ability to recognize objects and retrieve near-duplicate (to the query) images, this advantage over global features such as CCDs is diminished when visual diversity is enhanced by using a secondary medium, such as text, to pre-filter images. Thus, BOVW are not suitable for the proposed two-stage setup. In practice, applications like Google Goggles, where a user is querying an image in order to recognize a logo or a famous painting, BOVW models should be more effective. But in applications like Google Similar Images, where images are pre-filtered by text similarity, global features should be more suitable.

5. Two-stage vs. fusion of media

Fusion of different media into a single ranking is not trivial (van Leuken et al., 2009). In this section, we provide an experimental comparison of fusion to two-stage retrieval. Although in Section 2, we argued theoretically against fusion, in view also of the underlying assumption, hypothesis and drawbacks of two-stage retrieval, a comparison of the effectiveness of the two methods is in order.

Various techniques have been proposed to effectively fuse multimedia, e.g. simple linear weighting, principle component analysis (van Zwol, 2005), or using a weighted schema for aggregating features based on item scores (Wilkins, Ferguson, & Smeaton, 2006). In Zhou, Depeursinge, and Müller (2010), traditional approaches such as maximum combination (combMAX), sum combination (combSUM), and others, were employed. The results show that fused runs outperform the best single-medium runs. Several approaches for visual and textual information fusion that have been used in ImageCLEF over the past seven years are described in Mller et al. (2010). In Atrey, Hossain, El-Saddik, and Kankanhalli (2010), the authors provide an overview of the state of the art fusion strategies, which are used for combining multiple modalities in order to accomplish various multimedia analysis tasks.

5.1. Fusion of media

In Section 3.3 we elaborated on how to fuse image modalities. Here, we are extending this into fusing media, in a way suitable for the current setup. Starting from Eq. (1) and incorporating text as orthogonal medium, we add its contribution. Thus, the score s for a collection image against the topic is defined as:

Table 3
Retrieval effectiveness and matching time. The best results per measure and retrieval type are in boldface. Significance-tested with a bootstrap test, one-tailed, at significance levels 0.05 ($^{\triangle\triangledown}$), 0.01 ($^{\blacktriangle\triangledown}$), and 0.001 ($^{\blacktriangle\blacktriangledown}$), against the text-only baseline.

item scoring by	θ	MAP	P@10	P@20	bpref	Aver. Time (sec.)
text-only	-	.1293	.3614	.3307	.1809	-
$\max_i(\text{TOPSURF10k})_i$	.8000	.1158$^{\triangledown}$	.3786^{-}	.3286^{-}	.1700^{-}	.4359
	.5000	.1000$^{\blacktriangledown}$	.3243^{-}	.2871$^{\triangledown}$	.1582$^{\triangledown}$	.7748
	.3333	.0895$^{\blacktriangledown}$	.2729$^{\triangledown}$	.2579$^{\triangledown}$	.1513$^{\blacktriangledown}$	.7793
$\max_i(\text{TOPSURF200k})_i$	.8000	.1034$^{\blacktriangledown}$	.2886$^{\triangledown}$	.2600$^{\blacktriangledown}$	.1488$^{\blacktriangledown}$	.7714
	.5000	.1152$^{\triangledown}$	.3329^{-}	.3000^{-}	.1570$^{\triangledown}$	.6665
	.3333	.0959$^{\blacktriangledown}$	.2657$^{\triangledown}$	.2250$^{\blacktriangledown}$	.1395$^{\blacktriangledown}$	.4185
Equation 1	.8000	**.1428**$^{\triangle}$	**.4443**$^{\triangle}$	**.3857**$^{\triangle}$	**.1959**$^{\triangle}$	.0360
	.5000	.1405$^{\triangle}$	.4357$^{\triangle}$	.3821$^{\triangle}$	.1943$^{\triangle}$	.0432
	.3333	.1403$^{\triangle}$	.4357$^{\triangle}$	.3807$^{\triangle}$	.1942$^{\blacktriangle}$	.3632
$\max_i \text{JCD}_i$	.8000	.1348$^{\triangle}$	.4271$^{\triangle}$	.3743$^{\triangle}$	.1876^{-}	**.0095**
	.5000	.1305^{-}	.4171$^{\triangle}$	.3693$^{\triangle}$	.1859^{-}	.0110
	.3333	.1315^{-}	.4114^{-}	.3707^{-}	.1894^{-}	.2870
$\max_i \text{SpCD}_i$	.8000	.1342$^{\triangle}$	**.4443**$^{\triangle}$	.3771$^{\triangle}$	.1851^{-}	.0363
	.5000	.1302$^{\triangle}$	.4329$^{\triangle}$	.3693$^{\triangle}$	.1831^{-}	.0411
	.3333	.1307$^{\triangle}$	.4286$^{\triangle}$	.3743$^{\triangle}$	.1870^{-}	.0457

Please cite this article in press as: Arampatzis, A., et al. Dynamic two-stage image retrieval from large multimedia databases. *Information Processing and Management* (2012), http://dx.doi.org/10.1016/j.ipm.2012.03.005

$$s = (1 - w)\max_i \left(\sum_j \text{MinMax}(\text{DESC}_{ji}) \right) + w\ \text{MinMax}(\text{tf.idf}) \quad (2)$$

The parameter w controls the relative contribution of the two media; for $w = 1$ retrieval is based only on text while for $w = 0$ is based only on image. We report for five w values between 0 and 1.

5.2. An experiment

The general setup is similar to the one described in Section 3. For the two-stage experiment we threshold on prel with the SDTO. This was found in Section 3.5 to be more effective and robust than thresholding on estimated precision. We report for five prel thresholds.

Table 4 presents the effectiveness of fusion and two-stage against text- and image-only runs. Irrespective of measure, the best parameter values are roughly at: 0.6666–0.8000 for fusion's w, and 0.8000 for two-stage's θ. Both methods perform significantly better than text-only and far better than image-only. On the one hand, two-stage achieves better results than fusion, but it has more variability across topics: fusion passes the test at lower significance levels (i.e. higher confidence). On the other hand, effectiveness is less sensitive to the values of θ than the values of w: two-stage provides significant improvements in all measures for a wide range of thresholds (i.e. 0.3333–0.9900), while fusion can significantly deteriorate effectiveness for unsuitable choices of w.

Both methods are significantly better that text- and image-only baselines. Indicatively, the largest improvements in MAP against the text-only baseline are +9.0% and +10.4% for fusion and two-stage respectively, while the corresponding improvements in P@10 are +15.0% and +22.9%.

While two-stage performs better than fusion in 3 out of 4 measures, improvements are statistically non-significant at the 0.05 level. Further, both methods are robust in different ways: fusion provides less variability across topics but it is sensitive to the weighing parameter of the contributing media, while two-stage provides a much lower sensitivity to its thresholding parameter but has a higher variability. Nevertheless, two-stage has an obvious efficiency benefit over fusion: it cuts down greatly on costly image operations. Although we have not measured running times, only the 0.02–0.05% of the items (on average) had to be scored at the image stage. While there is some overhead for estimating thresholds, this offsets only a small part of the efficiency gains.

6. Related work

Image re-ranking can be performed using textual, e.g. (Kilinc & Alpkocak, 2009), or visual descriptions. Next, we will focus only on visual re-ranking. Subset re-ranking by visual content has been seen before, but mostly in different setups than the one we consider or for different purposes, e.g. result clustering or diversity. It is worth mentioning that all the previously proposed methods we review below used global image features to re-rank images.

For example, Barthel (2008) proposed an image retrieval system using keyword-based retrieval of images via their annotations, followed by clustering of the top-150 results returned by Google Images according to their visual similarity. Using the clusters, retrieved images were arranged in such a way that visually similar images are positioned close to each other. Although the method may have had a similar effect to ours, it was not evaluated against text-only or image-only baselines, and the impact of different values of K was not investigated. In van Leuken et al. (2009), the authors retrieved the top-50 results by text and then clustered the images in order to obtain a diverse ranking based on cluster representatives. The

Table 4
Retrieval effectiveness for fusion and dynamic two-stage retrieval. The best results per measure and retrieval type are in boldface. Significance-tested with a bootstrap test, one-tailed, at significance levels 0.05 ($^{\triangle\triangledown}$), 0.01 ($^{\blacktriangle\triangledown}$), and 0.001 ($^{\blacktriangle\blacktriangledown}$), against the text-only baseline.

		MAP	P@10	P@20	bpref
text-only		.1293	.3614	.3307	.1809
fusion w	.9000	.1380$^{\blacktriangle}$	.3786$^{\triangle}$	.3414$^{\triangle}$	.1901$^{\blacktriangle}$
	.8000	**.1410**$^{\blacktriangle}$	.4029$^{\blacktriangle}$	.3514$^{\blacktriangle}$	.1955$^{\blacktriangle}$
	.6666	.1403$^{\blacktriangle}$	.4129$^{\blacktriangle}$	**.3664**$^{\blacktriangle}$	**.1969**$^{\blacktriangle}$
	.5000	.1185 $^{-}$	**.4157**$^{\blacktriangle}$	.3657$^{\triangle}$	.1758 $^{-}$
	.3333	.0767$^{\blacktriangledown}$	.3871 $^{-}$	.3329 $^{-}$	.1278$^{\blacktriangledown}$
two-stage θ	.9900	.1376$^{\triangle}$	.4286$^{\blacktriangle}$	.3714$^{\triangle}$	.1899$^{\triangle}$
	.9500	.1390$^{\triangle}$	.4314$^{\blacktriangle}$	.3771$^{\triangle}$	.1917$^{\triangle}$
	.8000	**.1428**$^{\blacktriangle}$	**.4443**$^{\blacktriangle}$	**.3857**$^{\blacktriangle}$	**.1959**$^{\blacktriangle}$
	.5000	.1405$^{\triangle}$	.4357$^{\blacktriangle}$	.3821$^{\triangle}$	.1943$^{\blacktriangle}$
	.3333	.1403$^{\triangle}$	.4357$^{\triangle}$	.3807$^{\triangle}$	.1942$^{\triangle}$
image-only		.0107$^{\blacktriangledown}$	.0871$^{\blacktriangledown}$	.0871$^{\blacktriangledown}$	.0402$^{\blacktriangledown}$

Please cite this article in press as: Arampatzis, A., et al. Dynamic two-stage image retrieval from large multimedia databases. *Information Processing and Management* (2012), http://dx.doi.org/10.1016/j.ipm.2012.03.005

clusters were evaluated against manually-clustered results, and it was found that the proposed clustering methods tend to reproduce manual clustering in the majority of cases. The approach we have taken does not target to increasing diversity.

Another similar approach was proposed in Popescu et al. (2009), where the authors state that Web image retrieval by text queries is often noisy and employ image processing techniques in order to re-rank retrieved images. The re-ranking technique was based on the visual similarity between image search results and on their dissimilarity to an external contrastive class of diversified images. The basic idea is that an image will be relevant to the query, if it is visually similar to other query results and dissimilar to the external class. To determine the visual coherence of a class, they took the top 30% of retrieved images and computed the average number of neighbors to the external class. The effects of the re-ranking were analyzed via a user-study with 22 participants. Visual re-ranking seemed to be preferred over the plain keyword-based approach by a large majority of the users. Note that they did not use an image query but only a text one; in this respect, the setup we have considered differs in that image queries are central, and we do not require external information.

In Myoupo et al. (2009), the authors proposed also a two-stage image retrieval system with external information requirements: the first stage is text-based with automatic query expansion, whereas the second exploits the visual properties of the query to improve the results of the text search. In order to visually re-rank the top-1000 images, they employed a visual model (a set of images which depicts each topic) using Web images. To describe the visual content of the images, several methods using global or local features were employed. Experimental results demonstrated that visual re-ranking improves the retrieval performance significantly in MAP, P@10 and P@20. We have confirmed that visual re-ranking of top-ranked results improves early precision, though with a simpler setup without using external information.

Some other similar setups to the one we propose are these in Berber and Alpkocak (2009) and Maillot et al. (2006). In Berber and Alpkocak (2009), the authors trained their system to perform automatic re-ranking on all results returned by text retrieval. The re-ranking method considered several aspects of both document and query (e.g. generality of the textual features, color amount from the visual features). Improved results were obtained only when the training set had been derived from the database which is searched. Our method re-ranks the results using only visual features; it does not require training and can be applied to any database. In Maillot et al. (2006), the authors re-rank the top-K results retrieved by text using visual information. The rank thresholds of 60 and 300 were tried and both resulted to a decrease in mean average precision compared to the text-only baseline, with the 300 performing worse. Our experiments have confirmed their result: static thresholds degrade MAP. They did not report early precision figures.

7. Conclusions and directions for further research

We have experimented with two-stage image retrieval from a large multimedia database, by first using a text modality to rank the collection and then perform content-based image retrieval only on the top-K items. In view of previous literature, the biggest novelty of our method is that re-ranking is not applied to a preset number of top-K results, but K is calculated dynamically per query to optimize a predefined effectiveness measure. Additionally, the proposed method does not require any external information or training data.

The choice between static or dynamic nature of rank-thresholds has turned out to make the difference between failure and success of the two-stage setup. We have found that two-stage retrieval with dynamic thresholding is more effective and robust than static thresholding, practically insensitive to a wide range of reasonable choices for the measure under optimization, and beats significantly the text-only and several image-only baselines.

Additionally, we have compared fusion to dynamic two-stage retrieval. We have found that also fusion is significantly better than text- and image-only baselines. While two-stage performs better than fusion in 3 out of 4 measures, improvements are statistically non-significant at the 0.05 level. Further, both methods are robust in different ways: fusion provides less variability across topics but it is sensitive to the weighing parameter of the contributing media, while two-stage provides a much lower sensitivity to its thresholding parameter but has a higher variability.

A two-stage approach, irrespective of thresholding type, has also an obvious efficiency benefit over fusion: it cuts down greatly on expensive image operations. Although we have not measured running times, only the 0.02–0.05% of the items (on average) had to be scored at the expensive image stage for effective retrieval from the collection at hand. While for the dynamic method there is some overhead for estimating thresholds, this offsets only a small part of the efficiency gains.

We also investigated the performance of the visual codebook approach, specifically the TOP-SURF image descriptor, in the two-stage setup. We found that TOP-SURF is less effective than global descriptors such a CCDs: better than image-only but does not beat the text-only run. In efficiency, TOP-SURF is slower in matching speed than CCDs. Although the visual codebook methods are currently trendy because of their ability to recognize objects and retrieve near-duplicate (to the query) images, this advantage over global features is diminished when visual diversity is enhanced by first using a secondary text modality to pre-filter images. Thus, visual codebook methods are less suitable than global descriptors for the proposed two-stage setup.

There are a few interesting directions to pursue in the future. First, the idea can be generalized to *multi-stage* retrieval for multimodal databases, where rankings for modalities are successively being thresholded and re-ranked according to a modality hierarchy. In this respect, a suitable application may be SenseCam (Hodges, Williams, Berry, & Izadi, 2006)—the sensor-augmented wearable still camera. SenseCam captures a digital record of the wearer's day, by recording a series of images and other sensor data. The measurements taken from the sensors are used to augment and enhance the detection of concepts from visual features (Byrne, Doherty, Snoek, Jones, & Smeaton, 2010). In Doherty, Conaire, Blighe, Smeaton, and O'Connor (2008), the authors propose a late fusion method in order to combine information from image with data from

Please cite this article in press as: Arampatzis, A., et al. Dynamic two-stage image retrieval from large multimedia databases. *Information Processing and Management* (2012), http://dx.doi.org/10.1016/j.ipm.2012.03.005

the available sensors—such as Accelerometer, Temperature and Passive InfraRed (PIR)—in order to retrieve recorded events. In applications involving many media and huge amounts of data, multistage retrieval could be a method alternative to fusion, providing an effective and more efficient solution.

Second, we have not implemented any query classification to determine fusion weights dynamically per query (Yan, Yang, & Hauptmann, 2004) but used global weighting. This was a conscious decision in order to compare fusion to two-stage in their own merits without any training information. Recall that two-stage was also performed globally, by using a single threshold value for all queries optimizing some pre-defined effectiveness measure. In this respect, one may consider to implement query classification and investigate the relationship between query classes and suitable effectiveness measures to optimize.

Last, all experiments were performed on a standardized multimedia snapshot of Wikipedia; as an encyclopedic domain, this may be deemed narrow. However, our focus was to improve image retrieval with the help of other media beyond image, and this collection is one of the largest benchmark image databases for today's standards. It is also highly heterogeneous, containing color natural images, graphics, grayscale images, etc., in a variety of sizes, making it a challenging experimental setup. As more large multimedia testbeds become available, further experiments may reinforce our findings.

References

Aly, M., Welinder, P., Munich, M. E., & Perona, P. (2009). Automatic discovery of image families: Global vs. local features. In *ICIP* (pp. 777–780). IEEE.
Arampatzis, A., Kamps, J., & Robertson, S. (2009). Where to stop reading a ranked list: Threshold optimization using truncated score distributions. In *SIGIR* (pp. 524–531). ACM.
Arampatzis, A., Robertson, S., & Kamps, J. (2009). Score distributions in information retrieval. In *ICTIR. Lecture notes in computer science* (Vol. 5766, pp. 139–151). Springer.
Atrey, P. K., Hossain, M. A., El-Saddik, A., & Kankanhalli, M. S. (2010). Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems, 16*(6), 345–379.
Barthel, K. U. (2008). Improved image retrieval using automatic image sorting and semi-automatic generation of image semantics. *International workshop on image analysis for multimedia interactive services* (pp. 227–230).
Berber, T., & Alpkocak, A. (2009). DEU at ImageCLEFMed 2009: Evaluating re-ranking and integrated retrieval systems. In *CLEF working notes*.
Buckley, C., & Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. In *SIGIR* (pp. 25–32). ACM.
Byrne, D., Doherty, A. R., Snoek, C. G. M., Jones, G. J. F., & Smeaton, A. F. (2010). Everyday concept detection in visual lifelogs: Validation, relationships and trends. *Multimedia Tools and Applications, 49*(1), 119–144.
Chang, E., Goh, K., Sychay, G., & Wu, G. (2003). CBSA: Content-based soft annotation for multimodal image retrieval using bayes point machines. *IEEE Transactions on Circuits and Systems for Video Technology, 13*(1), 26–38.
Chatzichristofis, S. A., & Arampatzis, A. (2010). Late fusion of compact composite descriptors for retrieval from heterogeneous image databases. In *SIGIR* (pp. 825–826). ACM.
Chatzichristofis, S. A., Arampatzis, A., & Boutalis, Y. S. (2010). Investigating the behavior of compact composite descriptors in early fusion, late fusion, and distributed image retrieval. *Radioengineering, 19*(4), 725–733.
Chatzichristofis, S. A., Boutalis, Y. S., & Lux, M. (2010). SpCD—Spatial color distribution descriptor. A fuzzy rule based compact composite descriptor appropriate for hand drawn color sketches retrieval. In *ICAART* (pp. 58–63).
Chatzichristofis, S. A., Zagoris, K., Boutalis, Y. S., & Papamarkos, N. (2010). Accurate image retrieval based on compact composite descriptors and relevance feedback information. *International Journal of Pattern Recognition and Artificial Intelligence, 24*(2), 207–244.
Cula, O. G., & Dana, K. J. (2001). Compact representation of bidirectional texture functions. In *CVPR(1)* (pp. 1041–1047).
Doherty, A. R., Conaire, C. O., Blighe, M., Smeaton, A. F., & O'Connor, N. E. (2008). Combining image descriptors to effectively retrieve events from visual lifelogs. In *Multimedia information retrieval* (pp. 10–17).
Hodges, S., Williams, L., Berry, E., Izadi, S. Srinivasan, J. Butler, A., et al. (2006). Sensecam: A retrospective memory aid. In *Ubicomp* (pp. 177–193).
Kilinc, D., & Alpkocak, A. (2009). Deu at imageclef 2009 wikipediamm task: Experiments with expansion and reranking approaches. *Working notes of CLEF*.
Lewis, D. D. (1995). Evaluating and optimizing autonomous text classification systems. In *SIGIR* (pp. 246–254). ACM Press.
Li, X., Chen, L., Zhang, L., Lin, F., & Ma, W.-Y. (2006). Image annotation by large-scale content-based image retrieval. In *ACM multimedia* (pp. 607–610). ACM.
Li, J., & Wang, J. Z. (2008). Real-time computerized annotation of pictures. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 30*, 985–1002.
Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*(2), 91–110.
Maillot, N., Chevallet, J.-P., Lim, & J.-H. (2006). Inter-media pseudo-relevance feedback application to imageclef 2006 photo retrieval. In *CLEF working notes*.
Martinet, J., Chiaramella, Y., & Mulhem, P. (in press). A relational vector space model using an advanced weighting scheme for image retrieval. *Information Processing & Management*. doi:http://dx.doi.org/10.1016/j.ipm.2010.10.003.
Mikolajczyk, K., & Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(10), 1615–1630.
Mller, H., Clough, T., Deselaers, P., & Caputo, B. (Eds.). (2010). *ImageCLEF – Experimental evaluation in visual information retrieval*. Springer.
Mulhem, P., & Lim, J.-H. (2002). Symbolic photograph content-based retrieval. In *CIKM* (pp. 94–101). ACM.
Myoupo, D., Popescu, A., Le Borgne, H., & Moellic, P. (2009). Multimodal image retrieval over a large database. *Lecture Notes in Computer Science (Multilingual Information Access Evaluation II. Multimedia Experiments)*, 177–184.
Popescu, A. Tsikrika, T., & Kludas, J. (2010). Overview of the wikipedia retrieval task at imageclef 2010. In *CLEF (notebook papers/LABs/workshops)*.
Popescu, A., Moëllic, P.-A., Kanellos, I., & Landais, R. (2009). Lightweight web image reranking. In *ACM multimedia* (pp. 657–660). ACM.
Robertson, S. E., & Hull, D. A. (2000). The TREC-9 filtering track final report. In *TREC*.
Rodden, K., & Wood, K. (2003). How do people manage their digital photographs? In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 409–416). ACM.
Thomee, B., Bakker, E. M., & Lew, M. S. (2010). Top-surf: A visual words toolkit. In *ACM multimedia* (pp. 1473–1476).
van Leuken, R. H., Pueyo, L. G., Olivares, X., & van Zwol, R. (2009). Visual diversification of image search results. In *WWW* (pp. 341–350). ACM.
van Zwol, R. (2005). Multimedia strategies for B^3-SDR, based on principal component analysis. In *INEX. Lecture notes in computer science* (Vol. 3977, pp. 540–553). Springer.
Wilkins, P., Ferguson, P., & Smeaton, A. F. (2006). Using score distributions for query-time fusion in multimediaretrieval. In *Multimedia information retrieval* (pp. 51–60).
Yan, R., Yang, J., & Hauptmann, A. G. (2004). Learning query-class dependent weights in automatic video retrieval. In *Proceedings of the 12th annual ACM international conference on multimedia, MULTIMEDIA '04* (pp. 548–555). New York, NY, USA: ACM.
Zagoris, K. Arampatzis, A., & Chatzichristofis, S. A. (2010). www.mmretrieval.net: A multimodal search engine. In *SISAP* (pp. 117–118).
Zhang, S., Tian, Q., Hua, G., Huang, Q., & Li, S. (2009). Descriptive visual words and visual phrases for image applications. In *ACM multimedia* (pp. 75–84).
Zhou, X., Depeursinge, A., & Müller, H. (2010). Information fusion for combining visual and textual image retrieval. In *ICPR* (pp. 1590–1593).

Please cite this article in press as: Arampatzis, A., et al. Dynamic two-stage image retrieval from large multimedia databases. *Information Processing and Management* (2012), http://dx.doi.org/10.1016/j.ipm.2012.03.005