Iterative Relevance Feedback with Adaptive Exploration/Exploitation Trade-off

Nicolae Suditu and François Fleuret
Idiap Research Institute and École Polytechnique
Fédérale de Lausanne (EPFL), Switzerland

ABSTRACT

Content-based image retrieval systems have to cope with two different regimes: understanding broadly the categories of interest to the user, and refining the search in this or these categories to converge to specific images among them. Here, in contrast with other types of retrieval systems, these two regimes are of great importance since the search initialization is hardly optimal (i.e. the page-zero problem) and the relevance feedback must tolerate the semantic gap of the image's visual features.

We present a new approach that encompasses these two regimes, and infers from the user actions a seamless transition between them. Starting from a query-free approach meant to solve the page-zero problem, we propose an adaptive exploration/exploitation trade-off that transforms the original framework into a versatile retrieval framework with full searching capabilities. Our approach is compared to the state-of-the-art it extends by conducting user evaluations on a collection of 60,000 images from the ImageNet database.

Categories and Subject Descriptors

H.3.3 [**Information storage and retrieval**]: Information Search and Retrieval—*Search process, Relevance feedback*

General Terms

algorithms, experimentation, performance

Keywords

query-free interactive image retrieval, iterative relevance feedback, Bayesian framework, user-based evaluation

1. INTRODUCTION

It is recognized the need for image retrieval systems able to deal with automatically extracted content-based features, and provide an intuitive and simple interaction with users. A decade ago, the largest image collections were stock photograpy collections such as Getty Images and Corbis, containing hundreds of thousand images carefully annotated with keywords from a well specified vocabulary by experts with a homogeneous and professional knowledge. Nowadays, the on-line image collections such as Flickr or FaceBook are orders of magnitude larger. Although many images are annotated, the keywords are less reliable due to the users subjectivity and less consistent due to the uncontrolled vocabulary.

Research started to tackle this challenge via automatic tagging based on annotation propagation [13, 10, 18]. However, formulating a query might not be the most efficient way of searching for images since the visual content is often difficult to describe in terms of keywords. Relevance feedback is envisioned by many researchers as the only alternative that could cope properly with the challenges in image retrieval, and multimedia retrieval in general [12, 19].

The interactive retrieval process involves two different regimes. The first one can be seen as an *exploration* phase, during which the user communicates to the system her categories of interest in a broad way. This first regime transitions into the second one that can be seen as an *exploitation* phase, where the user specifies more detailed requirements on the visual properties of images, making the system intelligently explore the restricted subset specified during exploration.

We propose an extension of the retrieval approach developed by Ferecatu and Geman [6, 7] which has the major advantage of being query-free. Starting from an heuristic sampling of the collection, this method does not require any explicit query, and it relies solely on an iterative relevance feedback mechanism. At each iteration, the system displays a small set of images and the user chooses the image that best matches what she is looking for. The system updates an internal state and displays a new set of images accordingly. After a few iterations, the sets of displayed images start to include images that satisfy the user.

Our core contribution is an adaptive modulation of the exploration/exploitation trade-off, which leads to a versatile retrieval system with full searching capabilities. Internally, our approach employs an estimator of the consistency between the system internal state and the user retrieval objective, and controls dynamically, at each iteration, the selection of the displayed images accordingly.

We developed our system as a web-application, and we set it up for a collection of 60,000 images sampled uniformly from the ImageNet database [4], for which we took over the provided pre-computed SIFT features (Scale Invariant Fea-

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.
Copyright 2012 ACM 978-1-4503-1156-4/12/10 ...$15.00.

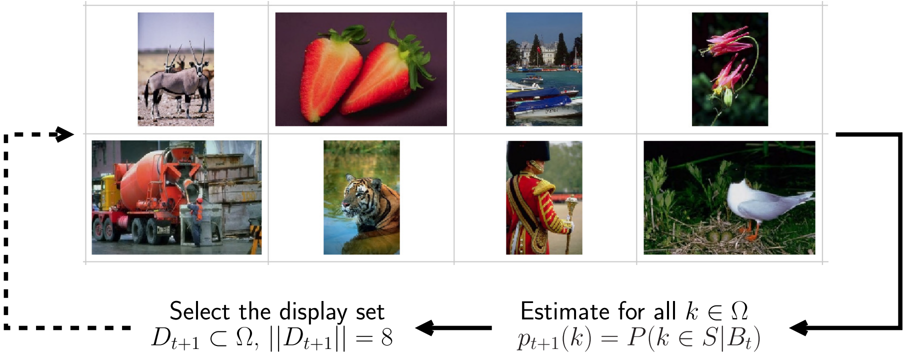

Select the display set
$D_{t+1} \subset \Omega, ||D_{t+1}|| = 8$
Estimate for all $k \in \Omega$
$p_{t+1}(k) = P(k \in S|B_t)$

Figure 1: Relevance feedback loop. At iteration t the system displays D_t. The next iteration $t+1$ is triggered by the relevance feedback event $\{D_t,\ x_t^*\}$. The system will update $p_{t+1}(k)$ for all $k \in \Omega$, and then it will select the new display set D_{t+1}.

ture Transform) [11]. We set up four configurations with different similarity metrics and we run user-based evaluations with 20 users. Evaluation gives evidence that our approach brings a significant improvement on the retrieval capabilities of the original system that remains sustainable when employing different similarity metrics.

This paper is structured as follows. In § 2 we present existing techniques related to the problem at hand, and summarize in § 3 the notation and the essence of the technique we are extending. In § 4 we elaborate our approach, and present in § 6 our experimental results. We conclude in § 7.

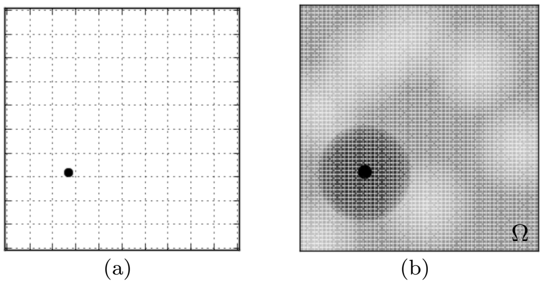
2. RELATED WORK

Research proposed many alternative approaches to tackle with the two retrieval regimes. Traditionally, they are seen as separate operations and they are treated by separate algorithms. Most of the image retrieval approaches require an initial query before offering relevance feedback tools. The most generic meanings are *query-by-visual-examples*[14], and *query-by-sketching*[8]. Regarding the relevance feedback tools, there are early works like MARS[1] and MindReader[9] that develope mechanisms for rich feedback information (*e.g.* ranking many images, tuning many parameters). Sharing the same line of thinking, a perception-based image retrieval system was developed by Chang et al. [2]. As reported in surveys [13, 17], there are many content-based image retrieval systems in research form but very few have been commercially developed.

The idea of searching images without any explicit query appeared distinctly in the work of Cox et al. [3]. The core of their work is a Bayesian framework for iterative relevance feedback. Ferecatu and Geman [6, 7] extended the framework and provided theoretically sound interpretations. Moreover, they conducted user evaluations that demonstrate the retrieval capabilities of such an approach. Their work focused on using a similarity metric based on low-level features extracted from the visual content (*i.e.* global descriptors of color, texture and shape). Recently, the framework was adopted and extended for large-scale image collections of millions of images in the HEAT retrieval system of Suditu and Fleuret [15]. Motivated by the potential of this query-free retrieval approach, our research takes a complementary direction and improves on its searching capabilities.

Table 1: Notation

Ω	image collection, where the images are identified by their indexes $\{1, 2, \ldots\ k, \ldots\}$
$S \subset \Omega$	set of images that the user is looking for
$D_t \subset \Omega$	set of images shown to the user at iteration t
$x_t^* \in D_t$	image chosen by the user at iteration t
$\{D_t,\ x_t^*\}$	relevance feedback event at iteration t
$p_t(k)$	probability of relevance of image k at iteration t
m_t	target mass for building the display set in the original system
m_t^{zoom}	target mass in the mass-zoom approach
z_t	change of the target mass at iteration t



(a) (b)

Figure 2: Abstract representation with a synthetic collection. (a): The images have as visual content one single point in the 2D Cartesian space, and the indexing features are the corresponding coordinates of that point. The similarity distances between images are the Euclidean distances between their corresponding points. (b): The duality between points and images is used to represent the entire collection. Additionally, the grey-levels of the points tell the probabilities of relevance of their corresponding images.

3. RETRIEVAL FRAMEWORK

This section presents briefly the retrieval framework proposed in [7]. Given a collection of images $\Omega = \{1, 2, \ldots k, \ldots\}$, the retrieval objective is to identify the small subset $S \subset \Omega$ containing all the images that the user is looking for.

The retrieval framework embodies an iterative relevance feedback mechanism that has two components. First, there

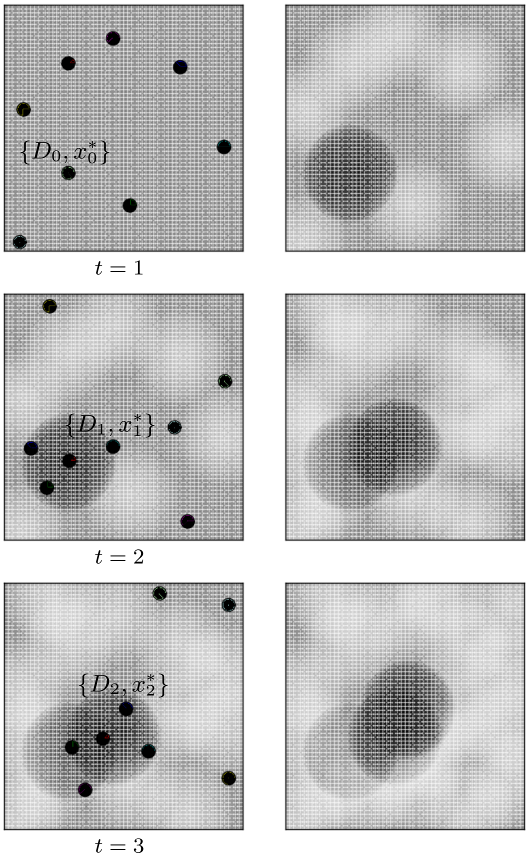


Figure 3: The posterior probabilities $p_t(k)$ for all $k \in \Omega$ are updated iteratively. Here, the relevance feedback events are given by a user who is searching for a point located in the center of the square. One can see how the distribution of probabilities evolves towards matching the user retrieval objective.

is a Bayesian framework that models the probabilities of relevance of the images in the collection as conditional probabilities depending on the relevance feedback events. Second, there is the strategy to select what images to show next given the estimates of the probabilities of relevance of all the images in the collection.

For an intuitive illustration of the system behavior, a synthetic image collection comes in handy, where the images have as visual content one single point in the 2D Cartesian space, and the indexing features are the corresponding coordinates of that point. Figure 2 explains the abstract representation based on the synthetic collection.

3.1 Posterior probabilities of relevance

Relevance feedback events are accumulated iteratively as shown in Figure 1. After the system displays a small set of images $D_t \subset \Omega$, $\|D_t\| = 8$, the user chooses one single image $x_t^* \in D_t$ that she considers to be the most similar to S, and this event is denoted as $\{D_t,\ x_t^*\}$. The cumulative event up to iteration t can be expressed as:

$$B_t = \cap_{i=0}^{t} \{D_i,\ x_i^*\} \quad \forall t \geq 0 \tag{1}$$

The conditional probabilities $p_{t+1}(k) = P(k \in S|B_t)$ are estimated after each relevance feedback event. Initially, when there is no relevance feedback yet, the probabilities $p_0(k)$ are initialized with 0.5 for all $k \in \Omega$. Subsequently, the conditional probabilities are estimated via an image similarity model defined over the metric space of the indexing features.

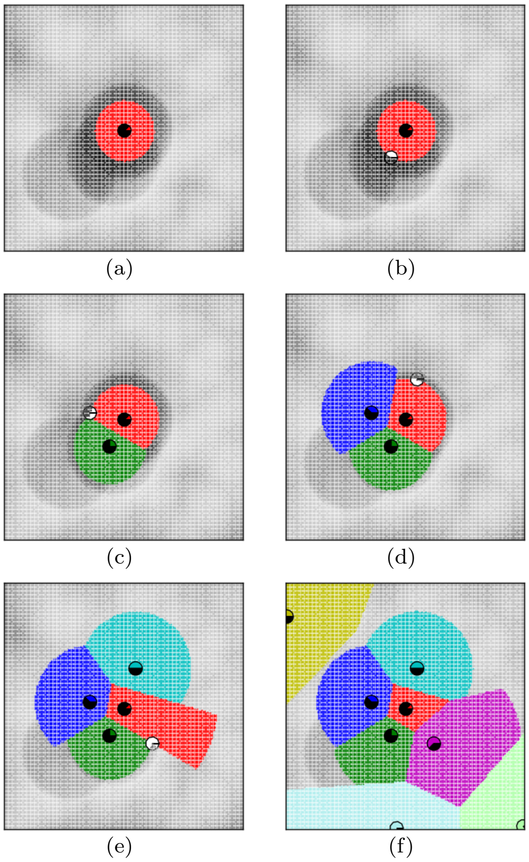

(a)
(b)
(c)
(d)
(e)
(f)

Figure 4: The set of displayed images is generated via the Voronoi tessellation algorithm. To illustrate its intermediate steps, the images already selected are marked in black and their current Voronoi cells are indicated by colors. (a): The first image x_0 is selected, and the first Voronoi cell $\mathcal{C}_0$ is grown. (b): The second image x_1 is selected. (c): The Voronoi cells $\mathcal{C}_0$ and $\mathcal{C}_1$ are grown in parallel. $\mathcal{C}_0$ is shrunken by detaching the images closer to x_1, and then regrown by including other images that are still closer to x_0. (d-f): The algorithm proceeds in the same manner until the set of displayed images is complete.

To simplify the analysis of our work, we use for that matter the exact same model as in [6, 7], which puts higher probability on the images similar to the chosen ones and accounts for an effect of “saturation” that ignores the increase in the image dissimilarities beyond a certain threshold. Figure 3 shows how the probabilities of relevance are gradually updated on successive iterations.

3.2 Selection of the displayed images

The displayed images, namely D_t with $\|D_t\| = 8$, are generated via a Voronoi tessellation algorithm proposed by Fang and Geman [5]. Instead of simply selecting the images with the highest probabilities of relevance, this algorithm samples the image collection with the purpose of maximizing the in-

Table 2: Both the baseline algorithm [7] and our mass-zoom method rely on the following procedures to compute a meaningful display set D_t. Given the current estimate of probabilities $\mathbf{p} = \{p_t(k)\ \forall k \in \Omega\}$, the cardinality $\|D_t\| = Q$, and a target mass m, the function ComputeDisplaySet returns a list of images $x_1, \dots, x_Q$ such that each of them has a high individual p_t, and they have disjoint neighborhoods $c_1, \dots, c_Q$ of mass m. Given the probabilities p, a list of images and a mass m, the function ComputeCells returns the corresponding disjoint neighborhoods, all of the same mass m.

```
Function ComputeDisplaySet($\mathbf{p}, Q, m$)
for $q = 1, \dots, Q$ do
  $c_1, \dots, c_{q-1} \leftarrow$ ComputeCells($\mathbf{p}, x_1, \dots, x_{q-1}, m$)
  $x_q \leftarrow \operatorname{argmax}_{k \in \Omega \setminus \cup_{i=1}^{q-1} c_i} p(k)$
end for
return $x_1, \dots, x_Q$

Function ComputeCells($\mathbf{p}, x_1, \dots, x_i, m$)
return $c_1, \dots, c_i$
  s.t. $\forall q\ \sum_{k \in c_q} p(k) = m$
  and $\forall q,\ r \neq q,\ \forall k \in c_q\quad \|k - x_q\| \leq \|k - x_r\|$
```

formation entropy, minimizing the redundancy between the displayed images, and thus maximizing the efficiency of the relevance feedback events.

The procedure **ComputeDisplaySet** to build a display set is described in Table 2. Given a target mass m, it picks each image successively, each time selecting the one with the highest p_t which does not belong to the neighborhoods of mass m centered on the images already selected. In the function **ComputeCells**, the neighborhoods are grown in parallel by including images one by one, as ordered by their similarity distances, until the probability mass of each neighborhood reaches the target mass m.

The original algorithm by Fang and Geman [5] uses at every iteration a target mass equal to a constant fraction of the total mass of the images

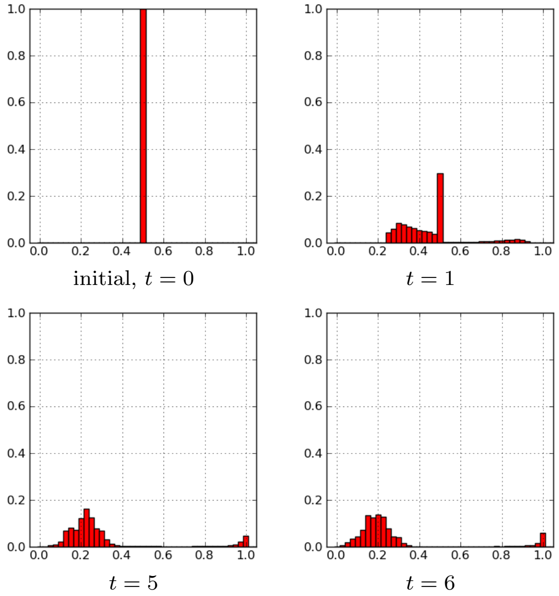
$$m_t = \frac{\sum_{k \in \Omega} p_t(k)}{\|D_t\|} \tag{2}$$

The first display set D_0 is generated by running the algorithm with the initial probabilities of relevance, $p_0(k) = 0.5$ for all $k \in \Omega$. The algorithm is still growing the Voronoi cells but it is choosing the images randomly between the equally probable candidates.

Figure 4 shows the intermediate steps of the Voronoi tessellation algorithm. One can see how the Voronoi cells are grown and how the images to be displayed are selected. Intuitively, the cells including regions with higher probabilities are smaller than the cells including regions with lower probabilities.

3.3 Limitations of the retrieval framework

As argued by Ferecatu and Geman [6, 7], the retrieval framework is well suited for image category search and that is, in our words, the first retrieval regime of exploring the image collection. They explicitly suggest that other retrieval techniques should be employed to retrieve specific images among these identified categories and that is, in our words, the second retrieval regime of exploiting the image collection.



initial, $t = 0$

$t = 1$

$t = 5$

$t = 6$

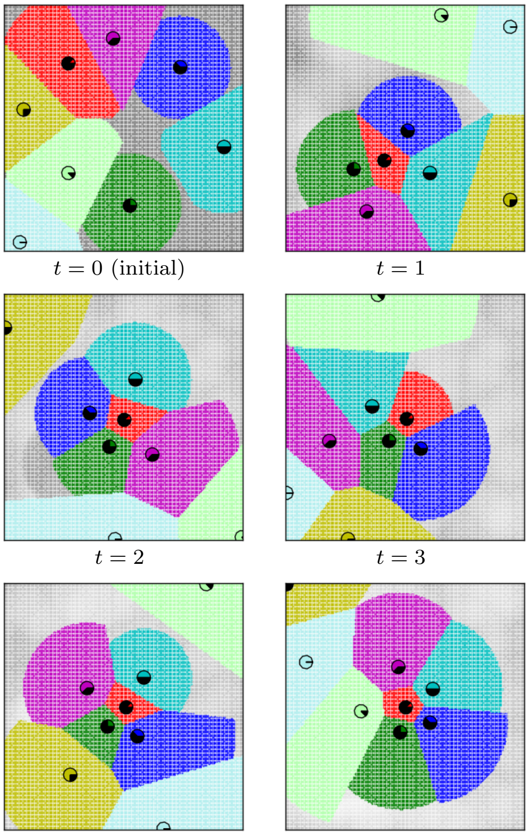
Figure 5: Evolution of the distribution of probabilities of relevance. The plots have the probability bins on axis X, and the percentage of images in the collection on axis Y. Initially, all images have the same probability, $p_0(k) = 0.5\ \forall k \in \Omega$. The distribution evolves rapidly in the first iterations, and it evolves slowly after the very first iterations.

A useful insight is given by analysing the evolution of a retrieval for the synthetic collection, when searching for a point located in the center of the square. Figure 6 shows the evolution of the displayed images, and Figure 5 shows the distribution of the probabilities of relevance.

As shown in Figure 5, the distribution of the probabilities evolves quite rapidly in the first iterations. These early iterations correspond to the first retrieval regime when the system is in the process of understanding broadly the categories of interest to the user. Later, after the system has achieved a good understanding of the user interest, the distribution of the probabilities evolves quite slowly from one iteration to another. These later iterations correspond to the second retrieval regime when the system is meant to refine the search and to converge to specific images.

As shown in Figure 6, the sets of displayed images include an image that is closer and closer, with each iteration, to the user interest. After 3 iterations, the system succeeds to display an image that is clearly in the intended region. Still after 5 iterations, the displayed images concentrate only slightly in the intended region.

The system succeeds efficiently to display an image in the intended region, but it has a hard time to display more and more images in the intended region. The “sampling” algorithm insists to cover the entire collection even after the distribution of probabilities becomes rather stable. One can say that the original system has a big inertia to maintain an exploration regime and it goes very slowly into an exploitation regime.



$t = 0$ (initial) $t = 1$

$t = 2$ $t = 3$

$t = 4$ $t = 5$

Figure 6: Evolution of the display set for the baseline algorithm with the synthetic collection, when searching for a point located in the center of the square. After 5 iterations, the displayed images concentrates slightly in the intended region. Again, the selected images are marked in black and their corresponding Voronoi cells are indicated by colors.

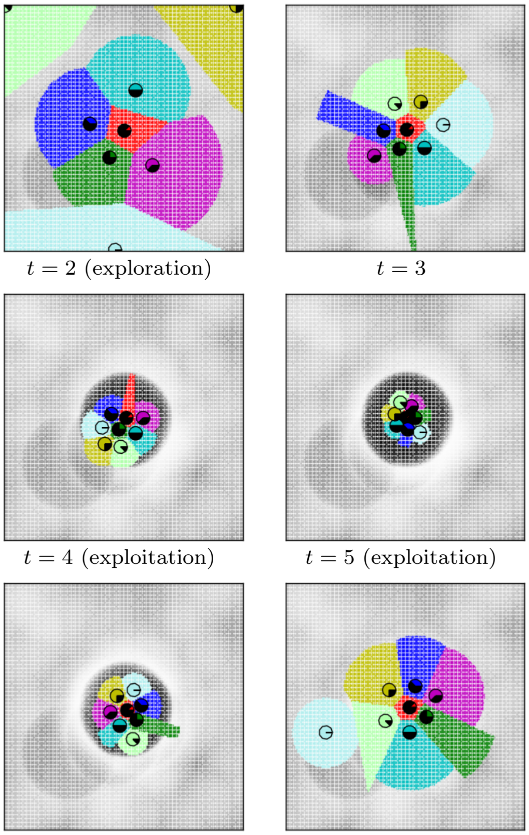
4. MASS-ZOOM SYSTEM

This section presents our solution to eliminate the limitations of the retrieval framework described in §3.3. Intuitively, the system should be aware of the degree of alignment of the distribution of probabilities with the user intent. When the distribution of probabilities is in line with the user intent, the system should to concentrate the "sampling" in the regions with high probability.

First, we present the sound idea of an adaptive strategy to handle the trade-off between exploration and exploitation, by modulating the concentration of the display set on promising images. Second, we present a heuristics that infers dynamically, at each iteration, from the user actions a consistency score that achieve a seamless trade-off that suits the user intent.

4.1 Exploration/exploitation trade-off

Our mass-zoom algorithm handles the trade-off between exploration and exploitation by modulating how much the display set should be concentrated on the images assessed as the most relevant. This is achieved by estimating at every iteration the target mass m_t for the displayed image neighborhoods (see § 3.2). While this value was a constant fraction of the total mass in the baseline (see Equation 2), we propose to link it to an estimate of the confidence of our current estimate of the image relevance. Making the value of this target mass smaller make the neighborhoods around the images of the display set smaller, which leads to a more compact display set, concentrated in the area of high probability.



$t = 2$ (exploration) $t = 3$

$t = 4$ (exploitation) $t = 5$ (exploitation)

$t = 6$ $t = 7$ (exploration)

Figure 7: Evolution of the display set for the mass-zoom system with the synthetic collection, when searching for a point located in the center of the square. After 5 iterations, the displayed images concentrates mostly in the intended region. The displayed images yet provide the freedom to escape the exploitation if necessary. The system continuously estimates the exploration/exploitation trade-off that suits the user.

Our approach increases the concentration of the display set if the choice of the user is consistent with our current estimate, and to decrease it otherwise. We propose the following update scheme:

$$m_t^{zoom} = z_t \cdot m_t \tag{3}$$

where $z_t \in \left(\frac{1}{m_t}, 1\right]$ accounts for the consistency between our estimates of the p_t and the user choice.

4.2 Heuristics based on a consistency score

The consistency score estimates the alignment of the system and the user intent. Immediately after the relevance feedback event $\{D_t,\ x_t^*\}$, the consistency score aims to estimate the alignment of the system and the user intent.

In the first iteration, the user intent is totally unknown and the consistency score c_0 is initialized to 1.0. Subsequently, the consistency score is estimated based on the probability of relevance of the chosen image $p_t(x_t^*)$ versus the probabilities of relevance of the other displayed images, namely $p_t(x_t)$, for all $x \in D_t$.

The consistency score is estimated based on the cumulative distribution function for the Gaussian distribution. The proposed heuristics is to scale up this value and to have a consistency score in the interval $[0.5,\ 2.0]$:

$$c_{t+1} = 0.5 + 1.5 \cdot \left(\frac{1}{2} + \operatorname{erf}\left(\frac{p_t(x^*) - \mu}{\sigma \cdot \sqrt{2}} \right) \right) \tag{4}$$

where

$$\mu = \frac{1}{\|D_t\|} \cdot \sum_{x \in D_t} p(x), \text{ and } \sigma^2 = \frac{1}{\|D_t\|} \cdot \sum_{x \in D_t} (p(x) - \mu)^2 \tag{5}$$

This is motivated by the intuition that if the $p_t(x_t^*)$ is already among the highest probabilities it means that the system has a distribution of the probabilities that is in line with the user intent, and thus the system is consistent with the user intent. If $p_t(x_t^*)$ is relatively low, the system is less consistent with the user intent.

The zoom value that impacts the exploration/exploitation trade-off of the selection of the displayed images is derived from the consistency scores as follows:

$$z_t = \prod_{i=0}^{t} \frac{1}{c_i} \tag{6}$$

4.3 Capabilities of the mass-zoom system

For an intuitive illustration, we run the mass-zoom system with the synthetic collection described in §3, and we search for a point located in the center of the square. We saved the evolution of the displayed images for intermediate iterations and we show them in Figure 7.

After efficiently identifying the intended region, the mass-zoom system is able to display more and more images in the intended region. The "sampling" algorithm concentrates in the intended region after the distribution of probabilities becomes rather stable. Although the "sampling" algorithm does not cover the entire collection anymore, the system continuously estimates the exploration/exploitation trade-off that suits the user.

Note that while the synthetic collection is very handy for intuitive illustrations, it should not be mistaken for image collections, which are typically facing high-dimensional image indexing feature spaces. Besides the miss-alignment between the image feature space and the user subjective perception of image similarities, the distribution of the image similarity distances impacts the Voronoi tesselation algorithm as well as the distribution of the probabilities of relevance. We argue that the exploration/exploitation trade-off has even higher impact than in the case of the synthetic collection.

5. SYSTEM OVERVIEW

The retrieval system was developed as a web-application (http://imr.idiap.ch/). Besides the advantage of permanent availability for evaluations, this implementation encourages the adherence to a realistic system architecture. The application software is distributed under the GPL v3.0 open-source license (http://www.idiap.ch/software/imr/).

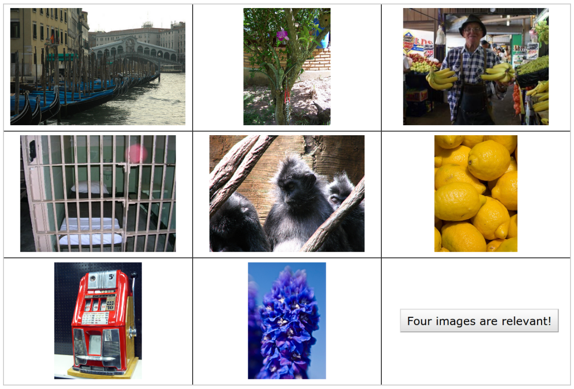


Figure 8: Web interface of the retrieval system used for user tests. The searching sessions were presented in a random fashion. The users were only told to end the searching sessions when they were satisfied by four of the displayed images.

The system was set up for 60,000 images sampled uniformly from the ImageNet database [4], that has the convenience of being structured in 1000 semantic categories, each composed of 500–2500 images. We considered the semantic information as benchmark meta-data for setting up the evaluation scenario, and we used as indexing features the pre-computed bags of SIFT features of dimension 1000 (Scale Invariant Feature Transform) [11], as they are provided along with the images. For evaluation purposes, we considered four different image similarity metrics defined over these histogram-like indexing feature vectors:

- Euclidean L^2 distance (L2)
- Isomap distance [16] derived from the 16 L^2 nearest neighbours (L2-Iso16)
- Manhattan L^1 distance (L1)
- Isomap distance derived from the 16 L^1 nearest neighbours (L1-Iso16)

The relevance feedback framework was calibrated as described in [6], and the parameters of the image similarity model are adjusted to saturate only after including on average 10% of the images in the collection. Therefore, each similarity metric would employ a different image similarity model, adapted to its statistical properties.

Computational effort required by the approach depends linearly on the collection size. Currently supporting multiple users, the web-application takes 1 second per iteration and uses 300KB cache memory per user with the data-set described in this article.

6. USER-BASED EVALUATION

Evaluation was conducted with 20 users not familiar with the system, and it consisted of running user tests with three systems: our proposed *zoom-mass* system, the original *baseline* system and a *random* system displaying images randomly without replacement. The random system discards totally the relevance feedback and thus provides the lowest possible performance.

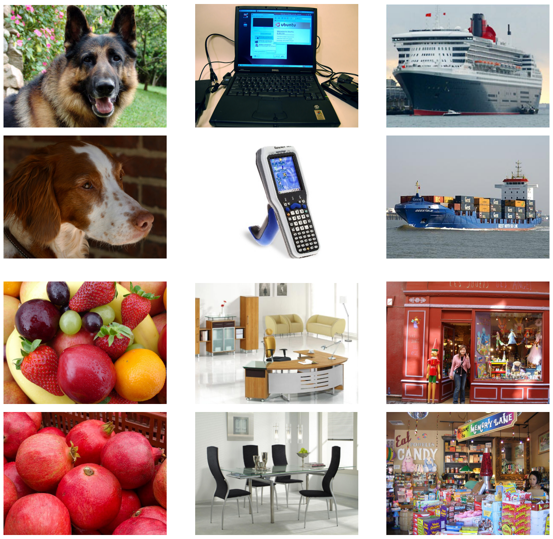


Figure 9: The users were asked to search for semantic categories described in words and accompanied by image examples as shown here. In order to ensure a sufficiently reliable diversity, there were 6 semantic categories.

6.1 Evaluation scenario

The aim of our experiments was to evaluate our *mass-zoom* system in terms of the retrieval capabilities, and to get evidence that our system is capable of providing capabilities beyond finding an image category, and is able to support refining the user interest in an efficient manner.

In order to isolate our contribution as much as possible, we employed four different similarity metrics on top of the image indexing features, as mentioned in §5. We did not aim to evaluate which similarity metric suits better the user subjective perception of image similarity, but rather to gather evidence that our contribution remains sustainable when employing different similarity metrics.

In order to ensure a reliable diversity, there were 6 semantic categories described in words and accompanied by the corresponding images in Figure 9:

- portraits/close-ups of dogs, wolves
- electronic devices as laptop, mobile phone
- big boats as ferryboats, cargoes
- baskets/plates with fruits, vegetables
- furniture items as tables, chairs
- entrances/windows of shops, shopping centers

In order to ensure comparable difficulty, these categories were chosen to be relevant for about 1% of our collection of 60,000 images based on the evidence given by the cardinality and the associated keywords of the ImageNet categories. Here, we should mention that these keywords were considered only as benchmark meta-data for assessing the retrieval difficulty, and our retrieval system does not make use of any textual information.

In order to avoid any bias, the searching sessions were presented in a random fashion. The semantic categories, the systems and the similarity metrics were randomized all together in one single user test. The users were not aware of which configuration was active in a certain session. In fact, they were not introduced to anything beyond the evaluation interface in Figure 8. The users were only told to end the searching sessions when they were satisfied by four displayed images.

We designed the evaluation scenario with the intent of pushing the evaluation beyond a simple image category search. We looked for evidence that the system is able to properly identify the user interest and then refine it more and more in an efficient way. This is the reason of asking the users to continue the searching sessions until four displayed images are relevant to what they search for.

6.2 Results analysis

Evaluation shows that the mass-zoom approach is viable. Mass-zoom is consistently better than the baseline for all configurations. Figure 10 shows the cumulative percentage of successful sessions per number of iterations. For example, for L1 similarity metric, *mass-zoom* finishes successfully in less than 10 iterations in 70% of the cases, and *baseline* in 45% of the cases. The *random* system is far from achieving the same performance even after 20 iterations. Table 3 contains a few discrete values read from Figure 10.

We argue that the system performs very reasonable when thinking of the most ideal case. If the collection would be arranged as a tree with 8 branches at each node, the *perfectly-structured* search will need around 3 iterations in average and $\log_8 N \approx 5$ iterations at maximum.

Table 4 tells about the statistical significance of the evaluation. For each couple of configurations, we counted how many times one performed better than the other for the same user and the same semantic category.Then, we computed the binomial probabilities. In principle, a difference is statistically significant if the corresponding probability is smaller than 0.05.

Figure 11 shows the evolution of the zoom values z_t from one iteration to the next. By decreasing in average, it shows that the system is consistent with the user interest. One should be aware that the rate by which the system transitions from exploration phase into exploitation phase (see Equation 4) may affect the results. An optimal rate could be derived by a more extensive user evaluation.

Metrics	Precision $(t < 5)$	Precision $(t < 10)$	Precision $(t < 15)$
L2	0.40/0.30	0.65/0.45	0.75/0.60
L2-Iso16	0.20/0.20	0.50/0.35	0.60/0.50
L1	0.30/0.25	0.70/0.45	0.80/0.65
L1-Iso16	0.20/0.15	0.40/0.20	0.50/0.30

Table 3: Retrieval performance. Here are a few discrete values read from Figure 10.

Metrics	Mass-zoom/Baseline	Mass-zoom/Random
L2	(35/60) 0.078	(53/60) 0.000
L2-Iso16	(40/60) 0.004	(50/60) 0.000
L1	(37/60) 0.026	(56/60) 0.000
L1-Iso16	(42/60) 0.001	(45/60) 0.000

Table 4: Binomial-test for statistical significance of our experiments for all four similarity metrics. For example, for L1 similarity metric, *mass-zoom* performed better than *baseline* in 37 times out of 60, and the probability of this to occur by chance is 0.026.

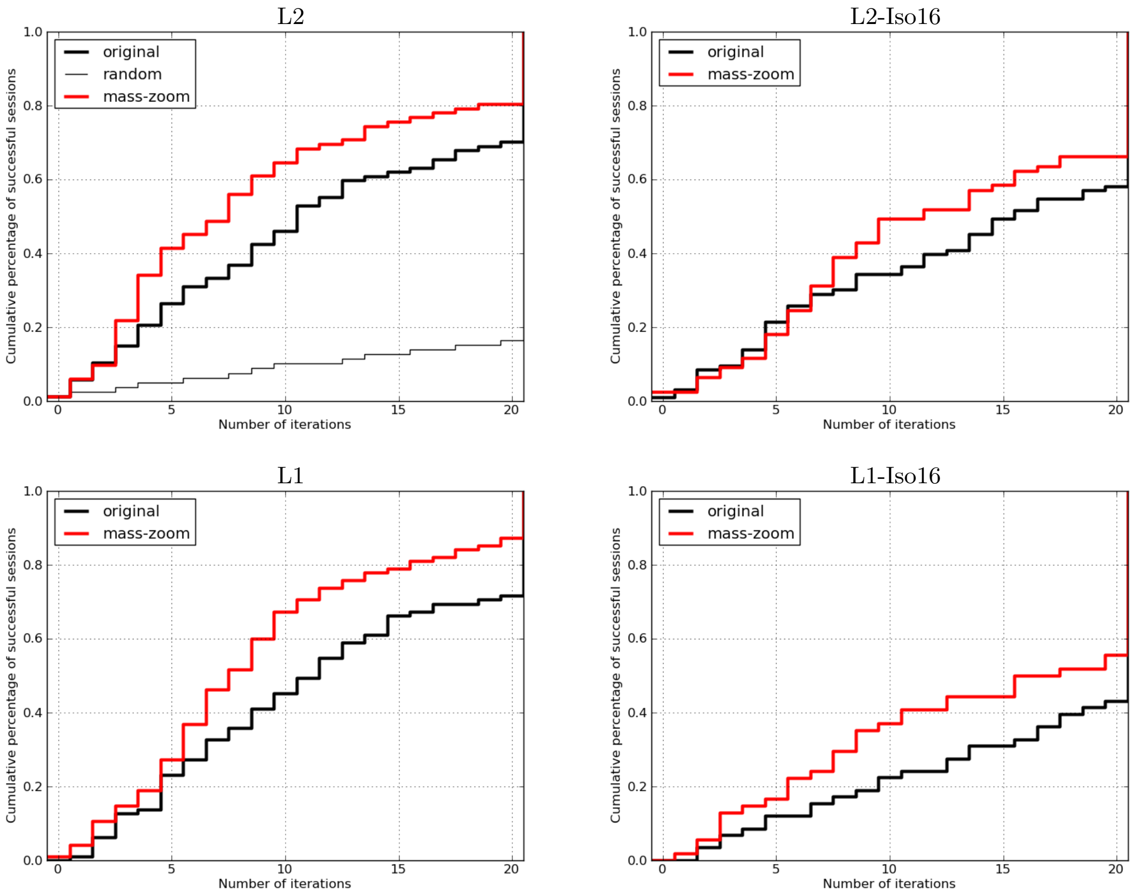


Figure 10: Cumulative percentage of successful sessions per number of iterations. Our *mass-zoom* system shows a sustainable performance against the *baseline* system proposed by Ferecatu and Geman [6, 7] for all four similarity metrics. For example, for L2 similarity metric, *mass-zoom* finishes successfully in less than 10 iterations in 65% of the cases, and *baseline* in 45% of the cases.

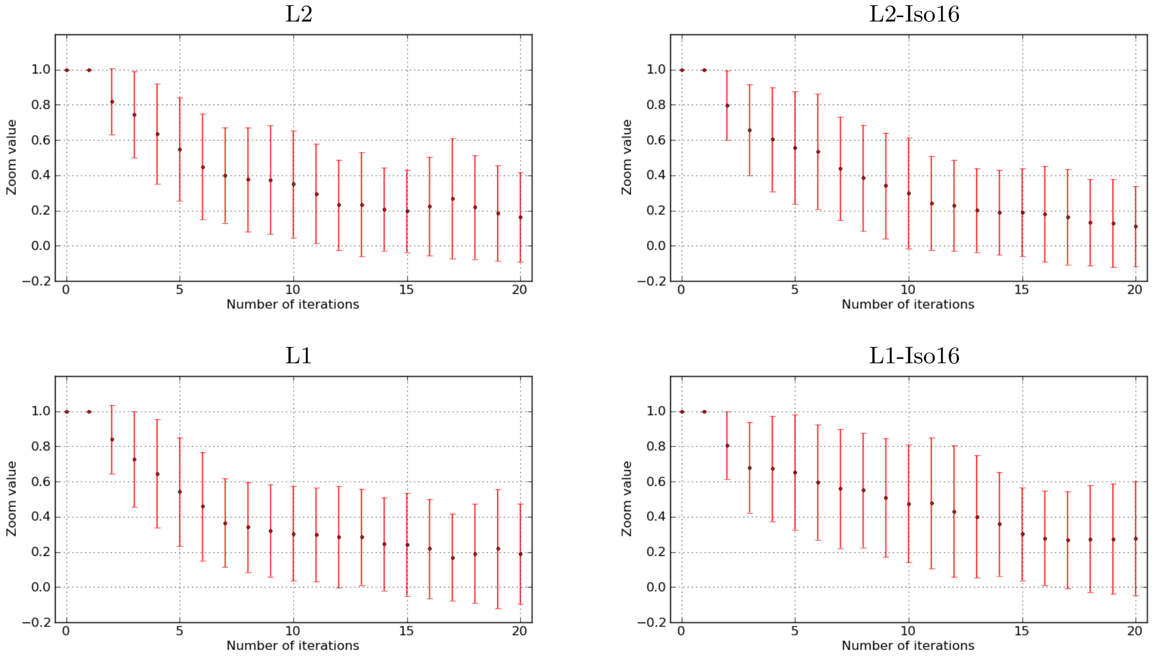


Figure 11: Zoom average and standard deviation. z_0 and z_1 are always equal to 1 as c_0 is initialized with 1 since there is no relevance feedback history at iteration $t = 0$, and c_1 is always equal to 1 since the probabilities of relevance $p_0(k)$ are all equal to 0.5.

Although we did not organize an appraisal questionnaire, we received favourable informal feedback regarding the user experience. The system is unconventional but intuitive and it becomes understood in very short time, even in the first searching session.

Suggestions have been made to improve the user experience. In the first couple of iterations, it may happen that none of the displayed images cannot be even vaguely related to what the user is searching for. When the users cannot make reliable similarity judgments, they would rather give *negative feedback* (*i.e.* none of the images resembles what they are searching for) or, at least, give *no feedback* and just ask for new images. Also, the users would appreciate the possibility to *undo* the last relevance feedback iteration. Such functionalities may be easily integrated in our approach, but they were intentionally not supported in the evaluation scenario.

7. CONCLUSION

We have presented a query-free retrieval approach with full searching capabilities. The adaptive mass-zoom system encompasses both retrieval regimes of exploration and exploitation, and supports a seamless transition between them that increases the alignment between the system and the user. Evaluation shows that our proposed mass-zoom system extends considerably the retrieval capabilities of the original algorithm. Moreover, evaluation gives evidence that the approach is intuitive and the minimalist user interface is effortless and self-explanatory.

The evaluation results give motivation for further investigations on how the system could benefit from other indexing features and similarity metrics. Although evaluated for image retrieval, our mass-zoom system is suitable for any type of multimedia retrieval with only minor changes.

8. ACKNOWLEDGMENTS

Nicolae Suditu was supported by the Hasler Foundation through the EMMA (Enhanced Medical Multimedia data Access) project. François Fleuret was supported in part by the European Community's Seventh Framework Programme FP7 - Challenge 2 - Cognitive Systems, Interaction, Robotics - under grant agreement No 247022 - MASH.

9. REFERENCES

[1] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, and J. Malik. BlobWorld: A system for region-based image indexing and retrieval. In *Proceedings of the 3th International Conference on Visual Information Systems*, volume 1614, page 660, January 1999.

[2] E. Chang, K.-T. Cheng, W.-C. Lai, C.-T. Wu, C. Chang, and Y.-L. Wu. PBIR: Perception-based image retrieval – A system that can quickly capture subjective image query concepts. In *Proceedings of the 9th ACM International Conference on Multimedia*, pages 611–614, October 2001.

[3] I. J. Cox, M. L. Miller, T. P. Minka, T. V. Papathomas, and P. N. Yianilos. The Bayesian image retrieval system, PicHunter: theory, implementation, and psychophysical experiments. *IEEE Transactions on Image Processing*, 9(1):20–37, January 2000.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[5] Y. Fang and D. Geman. Experiments in mental face retrieval. In *Proceedings of the 5th International Conference on Audio and Video-based Biometric Person Authentication*, pages 637–646, July 2005.

[6] M. Ferecatu and D. Geman. Interactive search for image categories by mental matching. In *Proceedings of the 11th IEEE International Conference on Computer Vision*, pages 1–8, October 2007.

[7] M. Ferecatu and D. Geman. A statistical framework for image category search from a mental picture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(6):1087–1101, June 2009.

[8] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. Query by image and video content: The QBIC system. In *IEEE Computer*, volume 28, pages 23–32, September 1995.

[9] Y. Ishikawa, R. Subramanya, and C. Faloutsos. MindReader: Querying databases through multiple examples. In *Proceedings of 24rd International Conference on Very Large Data Bases*, pages 218–227, August 1998.

[10] J. Li and J. Z. Wang. Real-time computerized annotation of pictures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(6):985–1002, 2008.

[11] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, November 2004.

[12] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Transactions on Circuits and Video Technology*, 8(5):644–655, 1998.

[13] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[14] J. R. Smith and S.-F. Chang. VisualSEEk: a fully automated content-based image query system. In *Proceedings of the 4th ACM international conference on Multimedia*, pages 87–98, 1996.

[15] N. Suditu and F. Fleuret. HEAT: Iterative relevance feedback with one million images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2118–2125, November 2011.

[16] J. Tenenbaum, V. de Silva, and J. Langford. A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science*, 290(5500):2319–2323, 2000.

[17] R. C. Veltkamp and M. Tanase. Content-based image retrieval systems: A survey. *Technical Report UU-CS-2000-34, Department of Computer Science, Utrect University, The Netherlands*, October 2000.

[18] J. Weston, S. Bengio, and N. Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. *Journal of Machine Learning*, 81(1):21–35, 2010.

[19] X. S. Zhou and T. S. Huang. Relevance feedback for image retrieval: A comprehensive review. *Journal of Multimedia Systems*, 8(6):536–544, 2003.