Media objects for user-centered similarity matching

Jean Martinet*, Shin'ichi Satoh
National Institute of Informatics
Tokyo, JAPAN
{jean,satoh}@nii.ac.jp

Yves Chiaramella, Philippe Mulhem
MRIM-CLIPS-IMAG
Grenoble, FRANCE
{First.Last}@imag.fr

Abstract

The increase of digital image and video acquisition devices, combined with the growth of the World Wide Web, requires the definition of user-relevant similarity matching methods providing meaningful access to documents searched by users among large amounts of data. The aim of our work is to define media objects for document description suited to images and videos, integrating a user-centered definition of importance for similarity matching. The importance is defined according to criteria and hypotheses, which have been experimentally validated. This leads to a definition of a weighting scheme for media objects (based on objects size, position, and scene homogeneity), which has also been validated with users in a second experiment. This model allows for meaningful similarity matching between document pairs and between users' queries and documents.

Keywords: Image and video indexing, Similarity matching, Human perception, Weighting scheme, Evaluation.

1 Introduction

Visual document retrieval has been an active research topic for more than 15 years now [39]. Many initial results have been achieved on content-based image and video retrieval systems considering mainly low-level features to represent the content of visual documents for retrieval purposes. Such approaches enable a user to search for documents according to a low-level description of the content. However, semantic descriptions of visual content are useful for a user, for instance in applications like home photographs management [29, 30]. Motivated by this reason, we consider visual content descriptions that contain semantics related to the visible objects of the displayed scene as well as to the overall organization of the image or video. This leads us to define so-called *media objects* that are generic semantic units used for content description, and we study more specifically the case of visual media objects.

Historically, classical models of information retrieval consider that a document is described by a set of representative index terms. In text retrieval for instance, index terms are keywords extracted from the collection. Because all index terms in a document do not describe it equally, they are assigned numerical weights. Popular weighting schemes for text retrieval are based on variations of *tf·idf* (*term frequency* × *inverted document frequency*) [2, 32, 33, 34]. The *tf* measures the importance of a term in a document, and is usually defined according to the (normalized) number of occurrences of the term in the document. The *idf* measures the discriminance of the term [35, 33] (also called *resolving power* [43]), i.e. the ability of the term to distinguish between the documents of the collection. The purpose of a weighting scheme is to give emphasis to important terms, quantifying how well they semantically describe and discriminate documents in a collection. The user's point of view is very important for defining weights for media objects in the context of information retrieval as such systems are designed for (and used by) humans [18]. In order to provide meaningful similarity matching for media objects between documents as well as between

*Jean Martinet is currently supported by the Japan Society for the Promotion of Science (JSPS). The authors wish to thank the helpful comments of the reviewers, which helped to improve the quality of this work.

users' queries, we also addressed the problem of assigning weights to media objects for the specific case of visual objects in order to represent their *level of interest* or *semantic importance* in a given document. For this purpose, we have identified several criteria related to the semantic importance of media objects from a user point of view. According to the identified criteria, we made corresponding hypotheses, which have been validated with users. The valid hypotheses drove the definition of an importance function for visual media objects. The most significant contributions of this paper are the definition of media objects as document descriptors, the proposition of the criteria and hypotheses about the importance, their validation with users, and the definition of a weighting model dedicated to similarity matching.

The remaining sections are organized as follows. In Section 2 we review the related work in similarity matching in images and videos. Section 3 describes the media objects and the specific case of image objects, and Section 4 defines the level of interest of image objects by identifying the criteria. Section 4 also describes the experimental validation of the hypotheses and defines the weighting model (in Subsection 4.4). Section 5 describes an experiment dedicated to evaluate the proposed weighting model in the context of a retrieval task. Conclusions and future work are discussed in Section 6.

2 Related work

Approaches in classical text retrieval are mostly based on keywords used as document descriptors [32]. The advantage of dealing with keywords is that the semantics is held by the keywords themselves, unlike in other media such as image or video. For instance, a text document in which the word *"boat"* appears is considered as a document about "boats". As a consequence, this document is considered relevant for the query *"boat"*. The indexing vocabulary in a text retrieval system is similar to the user's natural language expression. The words are automatically extracted, filtered, and weighted to give them more or less importance depending on how well they represent the document content.

When dealing with image and video documents the problem of the *semantic gap* [39] arises: because of the distance between the raw signal – that is to say the pixels – and its semantic interpretation, it is difficult to automatically extract an accurate semantic content representation of images and videos. Multimedia retrieval approaches usually consider low-level features (e.g. color, texture, etc.) as visual descriptors. Images and queries are represented with feature vectors, whose coordinates represent the amount of different colors or textures in the scene. Importance of image regions has been considered in several publications. In [23], Osberger and Maeder have defined a way to identify perceptually important regions in an image based on human visual attention and eye movement characteristics. In a similar way, Itti and Koch [8] have developed a visual attention system based on the early primate vision system for scene analysis. Later, Stentiford [41] applied the visual attention to similarity matching.

Region-based retrieval has been proposed in Blobworld [5] for instance. This system enables retrieving images according to low-level region similarity. Lu and Guo [16] have used color region clusters for image background identification and removal. Once the background regions are identified, they are removed from the image indexing process so that they do not interfere with meaningful image content. Wang et. al [47, 46] introduced in SIMPLYcity the *region frequency* and *inverse picture frequency* (*rf·ipf*), a region-based measure for image retrieval purposes, inspired from the *tf·idf* weighting scheme for text retrieval. The images in the collection are segmented into regions, employing several features (e.g., color components). The *rf* measures how frequently a region feature occurs in a picture, and the *ipf* attaches a higher value to region features that occur in fewer images, and that are therefore considered to be better discriminators. All regions are assigned an *rf·ipf* weight, which is stored for the image matching process. This weighting scheme has been evaluated for an image classification task with 10 categories such as *"beach"*, *"mountain"*, or *"building"*. Results show that a higher classification quality is achieved with this weighting scheme than when using simple color histograms. The basic difference is the integration of the regions distribution in the collection, which makes SIMPLYcity closer to semantic-based IR approaches. Later, Yahiaoui et al. [48] have used the *tf·idf* weighting scheme for video text summarization, as a parallel with text summarization techniques. Clusters of video keyframes are weighted according

to this formula in order to select the most appropriate clusters to build the video summary of video episodes. In the same time, Sivic and Zisserman [38] have also successfully used the *tf·idf* weighting scheme with visual keywords defined by quantizing visual descriptors into clusters. In their approach, the descriptors are elliptical affine invariant regions represented by 128-dimensional vectors using the SIFT descriptor developed by Lowe [14]. In our approach, we consider signal characteristics of image and video keyframe regions, as well as the overall document organization to define media objects and a corresponding weighting model.

This paper is an extended version of [18], which includes a generalization of the image object model to other media with extended criteria. The paper also presents a complete description of the image model, a more detailed analysis of the findings, and some additional experiments to validate the approach.

3 Media objects

What do users actually consider as significant when describing images and videos? How to provide a model reflecting their behavior about visual document relevance? According to previous experiments in the context of retrieving general domain images [7] and related to home photographs [12, 9, 22] users tend to describe images based on visible objects appearing in them, and users seem to find it more intuitive to access images using natural language descriptions, rather than low-level characteristics such as colors or textures [29, 30]. This observation motivates our approach focusing on "objects" and not on color regions for instance. It also explains why these objects are considered as basic entities in our model.

3.1 Definition

The definition of *media objects* (MOs) and specific instances of *video objects* and *image objects* is necessary in order to describe individual objects in multimedia documents:

Definition 1 (Media Object) *A media object is a semantic entity (from text, image, video, audio) in a multimedia document defined within spatial-temporal bounds.*

This definition is rather general. However, in this paper we study more specifically the case where Media Objects are visual objects coming from images or video streams, leading to the definition of *image objects* (IOs).

Definition 2 (Image Object) *An image object is a visual semantic unit in a multimedia document defined as a two-dimensional projection of one or more physical objects, which are parts of the scene displayed in an image or video.*

Image objects are a flexible model to describe image and video content, as – according to Definition 2 – an IO may correspond to several image or video keyframe regions, that are not necessarily spatially or temporally connected. Hence, they provide a solution for handling groups of objects in images, in agreement with the proximity principle of Gestalt perception of objects [13]. For instance, they can be used to describe a group of boats in a harbor, a group of trees in a forest, or a group of people in a crowd, at a given level of granularity. Furthermore, this flexibility makes it possible to take into account the frequent case where a physical object is spatially occluded (like a car behind a tree for instance), or temporally occluded from the scene in a video stream (like a person who is temporally hidden behind a passing car). We will further refer to these two situations as *fragmentation* cases of an IO.

3.2 Identification of media objects

In order to extract MOs from a multimedia document, they have to be identified by a segmentation process. For instance, for an image the space segmentation has to find a partition of the image by defining multiple regions identifying objects and boundaries. For an audio stream, the time segmentation generally categorizes the signal into *silence*, *noise*, *music*, or *speech*, and possibly

further refines the 3 last categories into the type of noise, the genre of music, or the identity of the speaker, respectively. The boundaries of these classes define the segmentation. For a video document, the space segmentation is the same as for image documents, and the time segmentation is obtained by detecting the shot boundaries.

We focus our study on IOs derived from images of video keyframes. The automatic segmentation and labeling of images and video keyframes is error-prone and remains an open problem in computer vision. However, the last 10 years have witnessed significant improvements in these tasks. The methods for segmentation include model-based segmentation [5], graph partitioning segmentation [37], clustering methods [47, 46], multi-scale segmentation [42]. The methods for annotation/recognition include neural networks classifiers [11, 12], SVM-based classification [44, 40, 1], for instance using popular SIFT features [15]. In order to avoid the imprecise segmentation process, an increasing number of approaches relies either on grid-based segmentations in which the images are tessellated in rectangular patches [11, 12, 47, 46, 4, 20], or on interest points detection [49, 28, 27, 6]. Local features extracted from patches or around interest points are then quantized and classified according to their visual appearance. In the grid-based approaches contiguous patches sharing the same label can be merged to form a larger region [12]. The annotation performances of automatic systems are steadily increasing: in [28, 27, 45], the authors achieve about 70% of annotation precision on a 6-class problem. In the case where the annotations are automatically derived, information such as the size and the position of corresponding regions can be directly used in the presented model.

3.3 Semantic interpretation of a media object

A media object is related to one and only one semantic interpretation which is defined by a *label*. A label stands here for a semantic interpretation of the related MO. In its general definition, a label may belong to any kind of indexing language (as part of a particular information retrieval indexing model). Considering general domain images and videos, such labels for IOs would be natural language words denoting real world entities (e.g. *"tree"*, *"house"*, *"sky"*, *"boat"*, etc.) Physical objects and IOs are connected through a semantic interpretation of MOs.

In the context of multimedia documents indexing and retrieval, according to the well known results achieved for text documents, a weighting method of the indexing elements can be defined to express their relative importance in the similarity matching process. Defining a *tf* equivalent for image and video documents is a challenging problem. We identify several criteria possibly having effects on the level of users' interest for IOs.

4 Level of interest

Having defined IOs, we now address the issue of assigning a *level of interest* (or *semantic importance*) to them, from a user perceptual point of view. In the context of visual document similarity matching, it is a central problem to assign weights to IOs in order to give emphasis to important IOs, quantifying how well they semantically describe documents. For this purpose, we discuss the *aboutness* of a document considering a given IO, that is to say the *importance* of the IO regarding the document. The challenge is to assign a weight to an IO in order to evaluate to what extent the multimedia document containing the IO is similar to another multimedia document, as well as to evaluate how relevant the document is regarding a user query about the IO. In our approach to image and video similarity matching, IO components are intended to provide a basis for both feature-based and semantic indexing of documents. Our definition of level of interest is related to the notion of document content in a much similar way to the standard notion of term importance in the context of text retrieval. The notion of occurrence – a valuable weighting element of index terms for textual documents – is much less intuitive in the case of multimedia documents. Images and video keyframes are two-dimensional data, and in this context one may postulate that the relevance of an image or a video keyframe showing some boats is not only related to the fact that it represents one or more distinct boats. Other two-dimensional perceptive (and possibly motion – when a sequence of keyframes is considered) factors are possibly more important in that matter, such as for example the overall scene organization of boats relatively to the background in the

scene. We describe in the following the criteria and hypotheses about the overall scene organization related to the importance of IOs, as well as two transversal factors to be taken into account when considering the criteria. However, variations of these two transversal factors are outside the scope of this study.

4.1 Criteria and hypotheses

Considering a document D and a visual media object mo (labeled with l) of D, the importance of mo regarding D is directly related to the aboutness of D considering the label l. For example, one will consider that an MO representing a boat (hence $l =$ *"boat"*) in a given video document is important if most users would consider that this video is about a *"boat"*. In our opinion, aboutness is a notion mainly local to a given document. It might not be confused with relevance, a notion that usually compares a particular document to an entire document collection. All the criteria presented here are local to documents and thus cannot be directly assimilated to relevance evaluation. A relation between these two notions is somewhat illustrated by the *tf·idf* weighting model in text retrieval: given a term t of document D, *tf* of t stands for the aboutness of D considering t, while *tf·idf* is an estimate of the relevance of D considering t [32].

Let us consider the proposed criteria and hypotheses about objects importance. Being associated to an image or keyframe region (or a set of regions), an IO is basically a set of pixels whose basic low-level and geometrical characteristics (color, texture, area, position, etc.) can be easily computed. Besides the geometry of regions, some other saliency aspects may possibly drive the IOs' importance – this is considered along with the hypothesis that there exists a relation between the saliency of a region and its semantic importance (for instance the eye saccade points follow a region sequence of decreasing semantic importance). According to the weighting models reported above in the related works and in Section 3 [32, 23, 5, 16, 39, 47, 46, 12, 22], we propose to investigate the following criteria.

4.1.1 Size

Some previous approaches have quoted the importance of the size criterion [47, 9], in which the authors assume a direct relationship between a media object importance and its normalized size (visible surface) S.

Hypothesis 1 (Size) *The importance of a visual media object varies in the same way as its size S.*

4.1.2 Position

Again, some previous approaches (e.g. Lim [12]) have quoted the potential importance of this criterion. We share this opinion and want to investigate possible formulations and their experimental validation. The position of an IO is determined by the position of its barycenter.

Hypothesis 2 (Position) *The importance of a visual media object is maximal when its position P is at the center of the scene, and decreases when its distance from the scene center increases.*

4.1.3 Fragmentation

To our knowledge, the possible impact of fragmentation on the importance of an IO has been investigated in [18] for the first time. We thought that it could indeed have such an impact, when compared to a completely connected, appearance of the same real object or that multiple occurrences of a given type of object could also have the same kind of impact. Fragmentation F of an IO is a more complex feature taking into account both the number and the size of the image or keyframe regions composing the IO.

Hypothesis 3 (Fragmentation) *The importance of a visual media object is maximal when it is not fragmented, and decreases when its fragmentation F increases.*

4.1.4 Homogeneity

Is the importance of any visible object in an image or a video keyframe depending on the fact that it appears alone (or almost alone) in the scene? Is this importance decreasing when the object appears among several others (of different nature) in the same scene? What we call the homogeneity criterion H of a scene expresses the extent to which a scene presents several types of IOs. To our knowledge, like the fragmentation criterion, this particular criterion was first considered in [18]. We suppose here that the more different IOs occur in a scene, the more the cognitive overload [1] of the user increases, and thus the more the detection of a particular IO becomes difficult. The homogeneity criterion H is then maximal when the scene contains only one type of IO and decreases when the number of occurrences of other types of IOs increases. The definition of H includes the number and the normalized surface of the various IOs in a given scene.

Hypothesis 4 (Homogeneity) *The importance of a media object varies in the same way as the homogeneity H of its embedding document.*

4.2 Motion

The definition of a weighting model for MOs requires to take into account the temporal structure of video contents. Indeed, object and camera motions can impact the users' perception of importance. For instance, a moving object draws one's attention more easily. As a consequence, all above listed criteria can be considered at a static level, that is to say just considering a still image configuration, or at a dynamic level, that is to say focusing on the possible change in the configuration.

Considering the above criteria, an object can become bigger or smaller, can move towards the center of the scene or towards the side, can become fragmented or aggregated (when corresponding regions gather or split), and the scene can become more or less homogeneous.

An example of the dynamic level for the size criterion is a scene in which a particular physical object gets closer to the acquisition device (hence the corresponding media object becomes bigger), possibly attracting one's attention. In this study, we choose to consider the static case of the criteria.

4.3 Context

Another transversal factor is the way criteria values are considered: absolutely or relatively to other IOs criteria values. The absolute interpretation consists in defining *a priori* the importance of IOs according to specific criteria values. The relative interpretation gives higher importance values to IOs whose criteria values significantly differ from other IOs criteria values.

For instance, considering the position criterion, an absolute interpretation would be to decide *a priori* that media objects are more important for a specific value of this criterion (for instance the center of the scene). A relative interpretation would correspond to assigning a high value whenever a specific object significantly differs from the remaining of the scene according to the given criteria, like an object alone far from the center, while all other objects are close to the center. In this study, we choose to consider the absolute interpretation for this criterion.

Together with the above identified criteria, the motion and context factors can be represented along three axis as shown in Figure 1. Of course, we do not claim that these criteria are the only ones related to an effective definition of the importance of an IO. We propose here a step for future evaluations of geometrical features of image and video objects on their importance. To our knowledge no similar study has been published about such a basic problem. For semantic image and video similarity matching, defining and validating such criteria is a first important step toward a well-founded weighting model for media objects. This experimentation process is described in the following section, considering the static case for motion and interpreting the criteria in an absolute context. Hence we evaluate the validity of hypotheses corresponding to the four criteria.

[1] Defined as "excessive demand made on the cognitive processes, in particular memory" in [25] page 717.

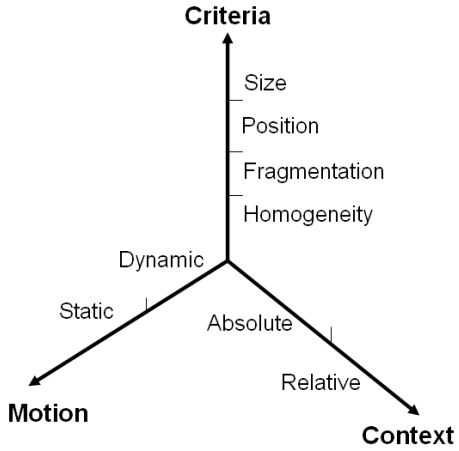


Figure 1: Criteria against motion and context factors.

4.4 Experimental validation of the hypotheses

The goal of this experimentation is to establish the validity of the hypotheses, by collecting user's assessment. For this purpose, we have designed an image test set, run an experiment with participants according to a protocol described below, and analyzed the data to draw statistical conclusions about the hypotheses.

4.4.1 Image test set

An image test set containing several configurations of image objects is designed. Relevant configurations have been determined according to the 4 hypotheses, considering 2 qualitative (absolute) values for each of them, which are **size**: big/small, **position**: center/lateral, **fragmentation**: aggregated/fragmented, **homogeneity**: homogeneous/heterogeneous. Considering all different combinations of these values, we get 16 ($=2^4$) configurations for an object (see Figure 2, where the IO is the disk – or group of disks).

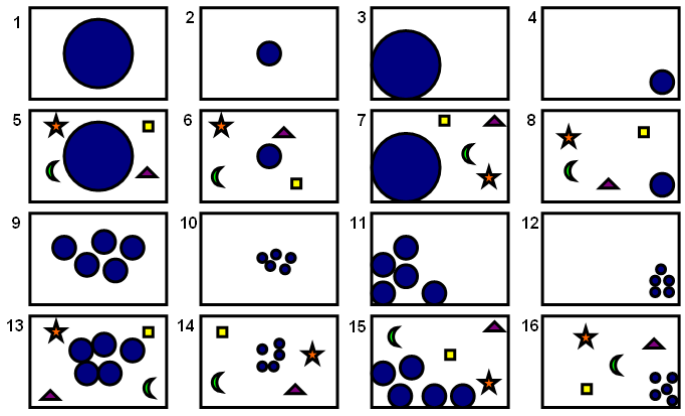


Figure 2: 16 typical configurations for an IO (the disk).

A way to prevent the results from being biased by the type of physical object, as well as to check whether the hypotheses hold across different kinds of physical objects, several object types covering a wide range of possible objects appearing in home photographs are selected. The study focuses on the following categories of objects:

- **Boat:** this category is used to reflect the fact that many non living objects occur in home photographs,

- **Bird:** we assume that animals often occur in home photographs, thus leading us to evaluate the criteria on such images,
- **Children face:** people faces are a very important part of home photographs, as underlined in [25]. In this test we used photographs showing on average as many girls as boys. For this object category and the following, we assume that many faces are present in home photographs.
- **One specific face:** in home photographs, many family of relative faces are present. During this experiment, the choice was made to use one of the authors face[2]. Note that the fragmentation criterion does not hold in this case as one specific person can not appear more than once in an image.

Figure 3 shows an example of two photos of a *boat* in different configurations: on the left, the boat is small, non-fragmented, in the center of a homogeneous image; on the right, the boat is small, non-fragmented, on the side of an heterogeneous image.

Figure 3: Example of photos showing configurations 2 (left) and 8 (right) of Figure 2.

Photographs have been carefully chosen, so that the following aspects remain as constant as possible in the collection: **the aesthetic quality** (images mainly come from a unique source, same non professional photographer), **the physical quality** (all images are clear, with luminosity and contrast corrected), **the image size** (all images are 400×300 in size), **the orientation** (all images are in *landscape* orientation), and **the typicality level** (objects for each category are as similar as possible; for instance, only white sailing boats are included in the boat category). Moreover, faces are full and looking at the photographer, and parity has been insured for the children faces. For each configuration and each object category, several images are available in the collection, and the collection contains 175 images.

4.4.2 The participants

The evaluation was carried out over 30 participants (14 men and 16 women), plus 5 participants for a pre-evaluation. The participants were educated people aged from 24 to 50, from our computer science research department. The consistency of the collected data has been verified during the evaluation, using the *split-half analysis consistency* method [17]. This method randomly splits in two sets X and Y the collected data, and computes the correlation value between the two sets X and Y. If the same conclusions can be drawn from the two sets X and Y, evaluated using the correlation between the sets X and Y, then the sample size is sufficient. For 30 participants, the correlation value is equal to 0.962, meaning that the collected data are highly consistent. Hence such a sample size is large enough, and makes it possible to perform statistical analysis over the collected data.

4.4.3 Protocol

The validation is based on the following protocol: participants are given the specific task to pick up the image out of two that is the most *representative* of (or *about*) a given object. The presented images show two of the typical object configurations (see Figure 2) from the same category that

[2]This choice is motivated by the fact that one author was present in the same room as the participants during experiments, and therefore his face is more easily recognizable for the participants than another arbitrarily chosen face. In the experiment section, this face is labeled "Jean".

are randomly selected from the collection. We proceeded on an exhaustive basis of all possible image pairs within each category. For the three first categories (boat, bird and children face) there are $C_{16}^2 = \frac{15\times16}{2} = 120$ configurations, and for the last category (specific face), there are $C_8^2 = \frac{7\times8}{2} = 28$, as the fragmentation criterion doesn't hold in this case. In total, $3\times120+28 = 388$ image pairs are presented to each participant. The average assessing time for one participant is 30 minutes.

4.4.4 Data analysis

The set of assessment data is an empiric distribution that can be modeled with probabilistic laws. We first analyze each criterion independently, then we analyze their combination.

Each of the 120 combinations involves 1, 2, 3 or 4 criteria. For instance, the pair (1,2) involves the size only, the other criteria remaining constant. The pair (6,16) involves the position and the fragmentation. The combinations can be grouped according the criteria they involve. These combination groups correspond to the elements of the set the parts of $\mathbb{C} = \{BS, CP, AO, HI\}$ where BS, CP, AO and HI (for *big size*, *center position*, *aggregated object* and *homogeneous image*, respectively) represent a criterion involved in a combination. Figure 4 shows the 15 combination groups[3], organized in a lattice. Each node indicates the involved criteria. For instance, the combination group $\{BS, CP, HI\}$ involves the size, the position and the homogeneity (the fragmentation remains constant); it corresponds to the 8 following combinations: (1,8), (2,7), (3,6), (4,5), (9,16), (10,15), (11,14) and (12,13).

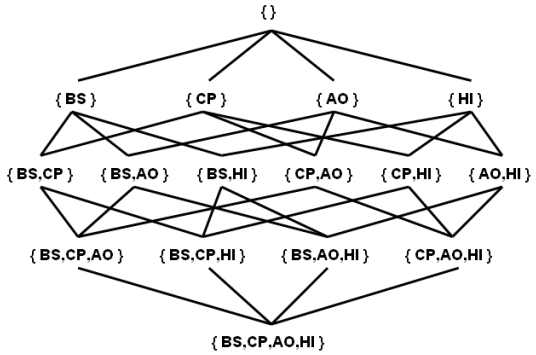


Figure 4: Combination groups.

When considering a pair (A,B) of images showing different configurations, we denote by V_A (resp. V_B) the criterion variation from B to A (resp. A to B), according to the hypotheses. Assume that the configuration for the IO in A is (big, center, aggregated, heterogeneous), meaning that the IO is big, located in the center of the image, aggregated, and that the image A is heterogeneous. Assume also that the configuration for an image B is (small, center, aggregated, heterogeneous). The image 5 (resp. 6) of Figure 2 is an example of A (resp. B). The criterion variation V_A is only based on the size S (the other criterion values remaining constant), big in image A and small in image B, denoted by (1, 0, 0, 0) according to the size hypothesis. Dually, the criterion variation V_B is denoted by (-1, 0, 0, 0).

We choose to model these data with a binomial law. The discrete random variable Y_e (where e is a combination group) is the Bernoulli variable, that can have values 1 or 0 with respective probabilities p_e and $1-p_e$. In the context of our evaluation, the experiment consists in presenting an image pair to a participant. The choice of an image is the result, which confirms or not one or several of our hypotheses:

$$Y_e = \begin{cases} 1 & \text{if the choice confirms the hypothesis of the group } e \text{ (case of success)} \\ 0 & \text{if the choice does not confirm the hypothesis of the group } e \text{ (case of non success)} \end{cases}$$

[3]We do not consider the empty set that corresponds to combinations involving no criterion.

If we associate to the possible values of Y_e the corresponding probability, we end up with the probability law Y_e. The expected value of Y_e is p_e, which is the probability to observe 1 as a result. The probability p_e is estimated over the sample according to the classical formula:

$$p_e = \frac{\text{number of favorable cases}}{\text{number of possible cases}}$$

We first focus on the 4 groups involving just one criterion (nodes of depth 1 in the lattice of Figure 4), from which we can decide about the validity of our hypotheses. We will then focus on the 11 other categories, corresponding to the criteria combinations (nodes of depth 2, 3 and 4 in the lattice of Figure 4). We need to estimate the probabilities of the binary random variables Y_e. The mean values and standard deviations of the results, averaged over the 30 participants are given in Table 1. For instance, the value 0.76 for the size criteria and boat represents the ratio of the number of choices confirming the size hypothesis over the total number of combinations involving the size only, for this object category. In other words, this value corresponds to the estimated probability that a participant finds an image of a big boat more representative of *boat* that an image of a small boat.

Criterion	**Boat**	**Bird**	**Child Face**	**Specific Face**
Size	0.76 (0.15)	0.74 (0.18)	0.77 (0.16)	0.97 (0.09)
Position	0.84 (0.13)	0.82 (0.12)	0.77 (0.15)	0.70 (0.27)
Fragmentation	0.38 (0.20)	0.53 (0.25)	0.60 (0.25)	–
Homogeneity	0.88 (0.12)	0.65 (0.19)	0.81 (0.12)	0.86 (0.22)

Table 1: Mean values (std.dev.) of the criteria for the 30 participants.

Table 1 shows that the fragmentation criteria seems to yield different results than the others: estimated probabilities for this criterion are close to 0.5, unlike the other criteria, whose probabilities are close to 0.8.

R.V.	**Proba. (std.dev.)**	**T-test (H_0)**
$Y_{\{BS\}}$	**0.81 (0.14)**	6.57E-13
$Y_{\{CP\}}$	**0.78 (0.16)**	1.84E-10
$Y_{\{AO\}}$	**0.50 (0.23)**	0.947954
$Y_{\{HI\}}$	**0.78 (0.14)**	6.74E-12

Table 2: Estimated probabilities (std.dev.) for all object categories, and T-test (Student) probabilities (H_0).

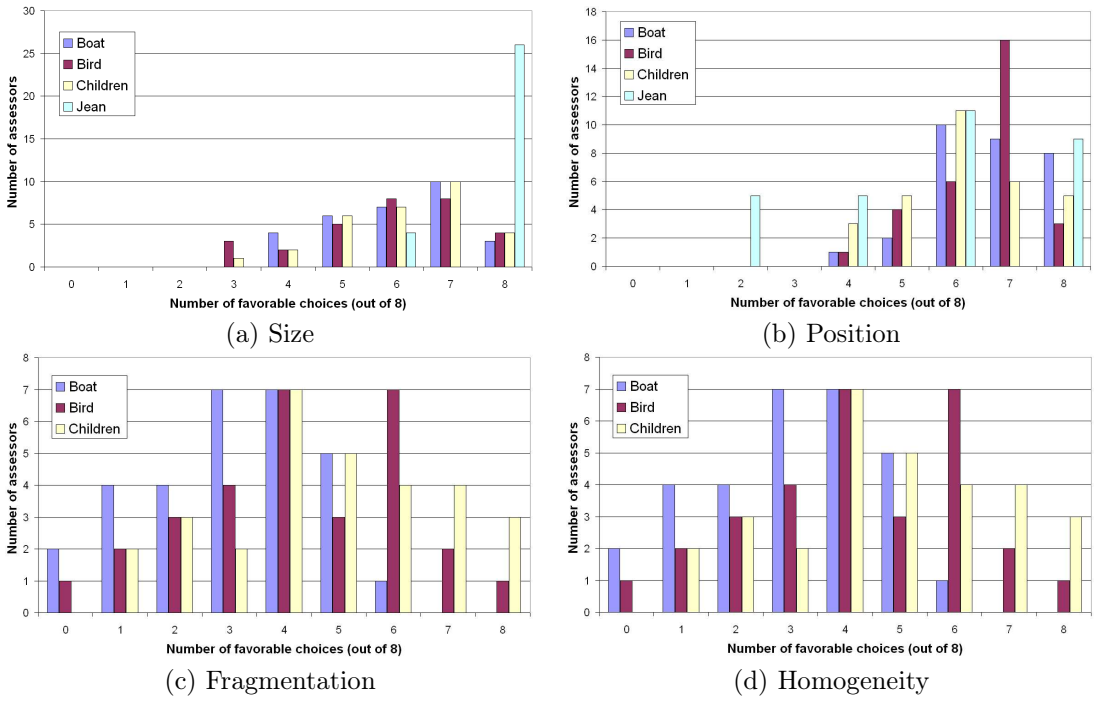
Estimated probabilities for random variables $Y_e, e \in \{\{BS\}, \{CP\}, \{AO\}, \{HI\}\}$ are presented in Table 2. Values range from 0.78 to 0.81, except the one corresponding to the fragmentation criteria, that is 0.5. A unilateral hypothesis test provides a statistical validation to our results. The null hypothesis is $H_0 : Y_e > 0.5$ for each criterion, and the alternative hypothesis is $H_1 : Y_e \leq 0.5$. Probabilities associated to the Student T-test [3] are also presented in Table 2. They correspond to the probability of wrongly rejecting the hypothesis H_0 while it's true. The test for the 3 criteria *size*, *position*, *homogeneity* are **very highly significant** with significance values much below the threshold of 1‰. Hence we can conclude that these 3 hypotheses are statistically valid. The T-test for the *fragmentation* gives a probability of 0.95, which doesn't allow us to conclude that this hypothesis is valid.

Size: The histogram of Figure 5(a) presents the empirical distribution of the participants' assessments for the criterion *size*. For instance, the bin 7 corresponds to the number of participants who selected 7 times out of 8 an image containing the big object, rejecting the image containing a small object.

One can notice that the distributions are quite similar through object categories: values are accumulated on the right-hand side of the distribution, which means that users' choices tend to favor bigger objects rather than smaller objects. It is even more obvious for the specific face, for

which the peak in the distribution indicates that a large majority of the participants has selected the image containing a big sized face (close-up or portrait). Indeed, humans and faces are usually the central subjects in images and videos. The first conclusion we can draw is that users tend to select images with big objects instead of the ones with small objects most of the times. This tendency is even stronger the the specific face. Hence, we can also conclude that different object categories may require different weighting methods. In this experiment, human faces are perceived by users in a different way from other objects.

We would like to highlight the fact that our study focuses on delimited physical objects with a concrete physical representation as a determined IO, which do neither e.g. include location references (such as "beach" or "Paris"), nor abstract notions (such as "vacation" or "happiness"), which have no direct and delimited physical representation in images or video keyframes.



(a) Size (b) Position

(c) Fragmentation (d) Homogeneity

Figure 5: Empirical distributions of assessors' choices according to the object category, for the criteria: (a) *size*, (b) *position*, (c) *fragmentation*. and (d) *homogeneity*.

Position: The histogram of Figure 5(b) presents the empirical distribution of assessment for the criteria *position*. The distribution is less homogeneous through object categories than the previous one, which means that this criteria doesn't have the same impact for all object categories. One can notice that the distribution for the specific face is more spread out than the others (which is confirmed by the relatively large standard deviation of 0.27 greater for the position of the specific face in Table 1). Indeed, 5 participants have selected only 2 times out of 8 the image containing the big specific face. The criterion *position* is less important in this case. We can conclude that humans and faces remains central regardless to their position in the scene.

Fragmentation: This criterion doesn't seem to be relevant for the modeling of IO importance. Statistical figures extracted from the data reflect the fact that, in average, participants have indicated **every second time** that the image containing the fragmented IO is more representative of the object than the image containing the non fragmented IO. This shows that in both configurations the images are found equally representative of the object. The histogram of Figure 5(c) presents the empirical distribution of assessment for the criteria *fragmentation*. As indicated by the estimated probability for this criteria, values are distributed around the central value 0.5. Moreover, distributions are similar for all object categories (we remind the reader that this criterion doesn't hold for the specific face).

Homogeneity: The histogram of Figure 5(d) presents the empirical distribution of assessment for the criteria *homogeneity*. This distribution is relatively homogeneous for all object categories.

To go further into the data analysis, we now focus on the criteria combinations, corresponding to the 11 lower categories in the combination group lattice in Figure 4. When 2 criteria are involved in a combination, the participant choice of a configuration may either confirm or deny both hypothesis, or confirm one hypothesis while denying the other. For instances, the choice of the configuration 1 from the combination (1,6) confirms the *size* and the *homogeneity* hypotheses, and the choice of the configuration 5 from the combination (2,5) confirms the *size* hypothesis and denies the *homogeneity* hypothesis. Table 3 presents the results using the following notation: the criterion whose hypothesis is denied by the test is noted in parenthesis. For instance, the variable $Y_{\{BS,(AO)\}}$ is bound to the probability associated to the choices confirming the *size* hypothesis and denying the *fragmentation* hypothesis. Note that the probability of $Y_{\{BS,(AO)\}}$ is the complement to 1 of $Y_{\{(BS),AO\}}$.

R.V.	Proba.	R.V.	Proba.
$Y_{\{BS,CP\}}$	**0.91**	$Y_{\{BS,(CP)\}}$	**0.54**
$Y_{\{BS,AO\}}$	**0.74**	$Y_{\{BS,(AO)\}}$	**0.70**
$Y_{\{BS,HI\}}$	**0.91**	$Y_{\{BS,(HI)\}}$	**0.47**
$Y_{\{CP,AO\}}$	**0.72**	$Y_{\{CP,(AO)\}}$	**0.73**
$Y_{\{CP,HI\}}$	**0.94**	$Y_{\{CP,(HI)\}}$	**0.51**
$Y_{\{AO,HI\}}$	**0.76**	$Y_{\{(AO),HI\}}$	**0.73**

Table 3: Estimated probabilities for the random variables involving 2 criteria.

In Table 3 we compare the results in each line. We see in the first line that the combination of the size and the position reinforces each of them taken independently: the probability for $Y_{\{BS,CP\}}$ is 0.91, which is greater than the probabilities for $Y_{\{BS\}}$ (0.76) and $Y_{\{CP\}}$ (0.81) alone (see Table 2). First line also shows that when the criteria have opposite variations, then the probability becomes close to 0.5. This means that, on average, the participants have selected indifferently a configuration with a big IO on the side of the image or a configuration with a small IO in the center of the image (configuration 3 of Figure 2 in the combination (2,3), for instance). This indicates that the criteria *size* and *position* are of equal impact regarding IO importance definition.

The other lines of Table 3 show that all relevant criterion pairs have the same behavior: their combination increases the probability, and their opposition makes it close to 0.5. Beside, it can be observed that the probabilities for the *size*, *position* and *homogeneity*, when combined with the *fragmentation*, do not change much when the *fragmentation* varies (see the values in lines 2, 4 and 6).

Table 4 shows the results for $Y_{\{BS,CP,AO,HI\}}$. The second column of this table contains the probabilities associated to this variable in the three following cases:

- the choice confirms all the hypotheses (first line),
- the choice confirms 3 hypotheses and denies the other hypothesis (four next lines),
- the choice confirms 2 hypotheses and denies the 2 other hypotheses (three last lines).

The results of Table 4 confirm the behavior of the variables found in Table 3: the combination of the valid criteria reinforces them, and the *fragmentation* does not have much impact.

4.5 Modeling

We have shown in the previous section that three out of four hypotheses are valid for the definition of IO importance: the ones concerning the size and the position of the IOs and the homogeneity of the images[4]. An important problem we are now faced with is the integration of these criteria within a model capable of reflecting the overall importance of an IO. We use the probability theory and

[4]Since the fragmentation criterion has been proved ineffective in the experiment (see Subsection 4.4), no formal modeling for F is given here.

R.V.	Probability
$Y_{\{BS,CP,AO,HI\}}$	**0.93**
$Y_{\{BS,CP,AO,(HI)\}}$	**0.80**
$Y_{\{BS,CP,(AO),HI\}}$	**0.91**
$Y_{\{BS,(CP),AO,HI\}}$	**0.86**
$Y_{\{(BS),CP,AO,HI\}}$	**0.83**
$Y_{\{BS,CP,(AO,HI)\}}$	**0.77**
$Y_{\{(BS),CP,(AO),HI\}}$	**0.69**
$Y_{\{BS,(CP,AO),HI\}}$	**0.71**

Table 4: Estimated probabilities for the variable $Y_{\{BS,CP,AO,HI\}}$ involving the 4 criteria.

Shannon information theory [36], which represents a formal framework suited for IO importance modeling. The aim is to associate an importance value for IOs based on each criterion value, where the criterion values are themselves related to IOs own geometric and position features. We first give a modeling of the three criteria, and then we show how they are combined into a single model of importance for IOs.

4.5.1 Size

Size of IOs is directly related to their surface, and consequently our size criterion is based on IOs normalized surface. Human perception of surfaces being rather logarithmic than linear [31], we define a normalized surface according to the following formula:

$$\mathrm{S}(o,I) = \frac{\log(n_p^o)}{\log(n_p^I)} = \log_{n_p^I}(n_p^o)$$

where o is an object, I is a given image, n_p^o is the number of pixels in o, n_p^I is the number of pixels in I, and $\log_{n_p^I}$ is the logarithm in base n_p^I. $S(o,I)$ is an increasing function of n_p^o (according to our first hypothesis) that has a logarithmic variation with values in $[\log_{n_p^I}(\epsilon), \log_{n_p^I}(n_p^I)]$, that is $[\log_{n_p^I}(\epsilon), 1]$, where ϵ is the minimum size (in number of pixels) of an object. Note that this interval becomes $[0,1]$ when $\epsilon = 1$.

4.5.2 Position

The position criterion P is integrated in our model through a non-uniform probability density function:

$$P(o,I) = p'_I(o)$$

where p'_I is a distribution of probability that gives higher probability values to IOs in the center of an image I. Some examples of such a distribution are given in Figure 6, where I is a simplified image seen as a unidimensional segment, and a two-dimensional distribution is shown above I. Two objects o_1 and o_2 are represented as parts of the segment, and their associated probabilities correspond to the black areas below the probability density curves.

The three distributions (b), (c) and (d) meet the requirements related to our second hypothesis, as the probability associated with o_1 is higher than the one associated with o_2. The position values belong to [0,1], and they are greater for IOs in the center of the image, according to our second hypothesis.

4.5.3 Homogeneity

With regard to the homogeneity criterion, let us consider normalized sizes of IOs as probabilities, as they fulfill all classical Kolmogorov properties of a probabilistic distribution and therefore can be treated as a probability. The probability associated to an IO o corresponds to the probability

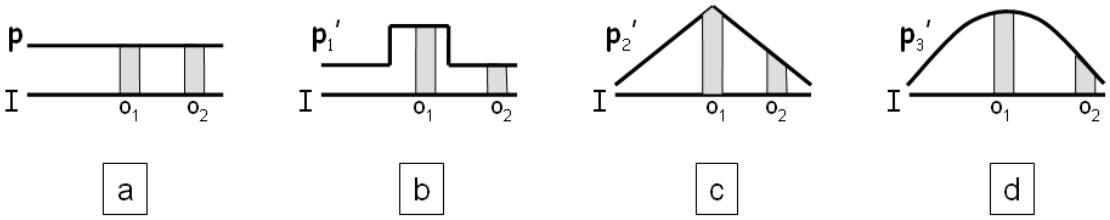


Figure 6: Uniform distribution of probability (a), and examples of non-uniform distributions of probability (b, c and d).

that a pixel taken randomly from the image belongs to o:

$$p(o, I) = \frac{n_p^o}{n_p^I}$$

Based on this probability, we can define a spatial entropy SH (according to Shannon [36]) computed from the spatial distribution of IOs in an image, in order to represent the "*disorder*" in the image:

$$SH(I) = -\sum_{o \subseteq I} p(o, I) \cdot \log\left(p(o, I)\right)$$

The term "$-\log\left(p(o, I)\right)$" represents the amount of information held by o in the context of I. Since IOs have a minimal size, $p(o, I) \in [\frac{\epsilon}{n_p^o}, 1]$, and consequently: $-\log\left(p(o, I)\right) \in [0, \log(\frac{n_p^o}{\epsilon})]$. Entropy values are larger when there are many objects with the same size in the image, i.e. when one object in particular is less easily visible. Entropy values are smaller when there are few objects with different sizes, and when one object is bigger than the others and therefore more easily visible in the image. To be consistent with our last hypothesis, the homogeneity criterion is defined as the complement to 1 of SH to which we apply a normalization factor. The homogeneity H is defined according to the following formula:

$$H(I) = 1 - \frac{SH(I)}{SHmax(I)}$$

where SHmax(I) is the maximum value of the spatial entropy, corresponding to the virtual case where the image is composed of n_p^I 1-pixel IOs. The homogeneity values are large for homogeneous images according to our last hypothesis, and they belong to $[0, 1]$. This value is the same for all IOs in one image.

4.5.4 Criteria combination

We now describe how the three criteria are combined into a single importance model for IOs. The importance value of an IO should be large when all criterion values are large, and inversely. Moreover, the importance value should not be close to zero when one of the criterion values is close to zero while the two others are large (see Table 2). For these reasons, a good candidate for the combination is the addition. Since the three criteria might have different variation ranges on different collections, a normalization scheme is applied in order to align both lower and upper bounds [10]:

$$v_n = \frac{v - min\_value}{max\_value - min\_value}$$

where min_value (resp. max_value) is the minimum value (resp. the maximum value) of the criterion value found the whole collection. We now define the importance of an IO o in an image I as:

$$\text{importance}_I(o) = S_n(o, I) + P_n(o, I) + H_n(I) \quad (1)$$

where S_n, P_n and H_n are the normalized values of criteria S, P and H respectively. Hence, our definition of IO importance combines the three criteria that have been experimentally validated to reflect the aboutness of images with regard to the semantic interpretation of IOs.

4.6 Summary and discussion

The section 4 provides a definition of the level of interest for IOs. IOs are a particular case of MOs, that can be used to model visual – image and video – data. Hence the matter discussed in this section is directly applicable to these two types of media. We have first identified four criteria and hypotheses likely to affect the semantic importance of IOs, that are the **size**, the **position**, and the **fragmentation** of the objects, as well as the **homogeneity** of the images in which IOs appear. We have also discussed about two transversal factors likely to impact the semantic importance of IOs, which however are not in the focus of our work. We have then presented an experimental evaluation for the criteria and hypotheses.

4.6.1 Findings

The experimental results indicate that 3 out of 4 hypotheses are statistically valid. We emphasize the two following aspects of this part of our work:

1. We provide a user validation of the criteria of **size** and **position**, which have been used in the past, without any solid evidence of their validity.

2. To our knowledge, the criteria of **fragmentation** and **homogeneity** have never been proposed nor investigated before our previous work presented in [18]. We identify these two criteria, and also provide a validation for the homogeneity criteria, as well as an invalidation for the fragmentation criteria, which is an important result of our study.

One other interesting outcome of the study is the universal agreement of the assessors for the importance of the size criteria for the specific face. This could be a starting point for a more refined study for human faces, that could result in a specific weighting method for human faces (which usually are the objects of interest in visual documents) that would give more importance to the size than to other criteria.

Another issue is related to the size of natural elements that are usually part of the background in images, and that are likely appear as IOs of big size, such as *"sky"*, *"sea"*, *"ground"*, or *"forest"*. One could argue that such elements should not be given a large importance since they are part of the background of the images and therefore they are less important than other objects, possibly smaller, such as a face. However, our weighting model will generally give a high importance to big objects. First, giving a large importance to a big "sky" is consistent with the fact that the considered image displaying a large part of sky is relevant to a query about *"sky"*, since it represents an answer to the need of the user. Second, our model deals with the *local* importance of IOs, which is different from its *global* discriminance value regarding the collection. If this model were to be used in a search system, it should be used together with a measure of the global importance of the term in the collection, such as the *idf* [35, 33]. If the collection contains mostly outdoor images with many "skies", then the global importance of *"sky"* will be low, and the overall importance given to *"sky"* will be decreased accordingly. We remind the reader that in the presented hypotheses, we want to investigate the *criteria* axis of Figure 1, and we choose to consider the static case for *motion* and to interpret the proposed criteria in an *absolute context*.

4.6.2 Analogy to text retrieval

We believe that considering labels associated with IOs of an image is a reasonable starting point to study its semantic content. Simple text retrieval approaches enable a user to retrieve documents according to the **keywords** that occur in the text. For instance, text documents including *"boat"* can be retrieved by a query containing this word. However, documents containing *"water-craft"* or *"canoe"* will not be retrieved if they do not contain the query term. In order to retrieve documents with content **related** to *"boat"* (e.g containing synonyms, hypernyms or hyponyms), it is necessary to include to the model a relation processing scheme, such as document/query expansion based on a thesaurus or an ontology [26, 21, 19]. The importance model presented here provides a basis for determining a counterpart of the textual *tf* for visual data. Therefore, it is meant to be used in a system together with such a relation processing module, that would – for instance – derive the importance of *"beach"* from the IOs *"sea"* and *"sand"*.

5 Experiments

The experiment described here is aimed at validating the criteria modeling and combination proposed in the previous section. From a test collection containing about 800 images and a set of 20 queries, we assess the ability of our model to retrieve and rank images according to the importance of objects. First we describe the image collection, and we report an experiment showing how the weighting scheme impacts the distribution of documents in the space. Then we show the results of a standard information retrieval evaluation on the collection with the queries. This evaluation compares our model with a Boolean weighting scheme, and with a content-based image search system based on color histograms. These retrieval evaluations are performed with a *Query By Example* search scheme. We finally present a last experiment specifically comparing the documents ranking of our system with a ground truth generated from users' aboutness assessments, in order to estimate how close to the users' perception the proposed model is.

5.1 Image test collection

For the purpose of evaluating our model, we have built an image test collection. This test collection consists of over 800 personal photographs taken by the authors. Images have been thoroughly segmented and indexed manually by authors[5]. The collection contains a wide range of authors' holidays photographs. Importance values of image objects are computed according to the equation 1.

We have also designed a set of 20 queries, in which each query consists of a single image from the collection displaying a target object (e.g. *"Horse"*, *"Pond"*, *"Foggy_ sky"*, *"Bridge"*, *"American_ Flag"*, etc.) These target objects correspond to different kinds of objects (natural objects like *"River"*) and non-natural ones like *"Bridge"*), and they also correspond to different levels of genericity of the objects (specific like *"American flag"* and generic like *"People"*).

5.2 Density of document space

Internal document representation determines the document distribution in the document space. Sparse document spaces are considered good for an information retrieval task, since a relevant document may be retrieved without necessarily retrieving many documents in its surrounding, which yields a good output precision [35]. Table 5 shows document space density values computed for the collection, according to the following formula:

$$density = \frac{1}{n} \sum_{i=1}^{n_{doc}} sim(d_i, d_k)$$

where d_i is the vector of importance values of IOs for a document i, d_k is the centroid vector of all documents, $sim(d_i, d_k) = \cos(d_i, d_k)$, and n_{doc} is the number of documents in the collection. This value corresponds to a average similarity between all documents and their centroid. A low density value indicates that documents are well separated. The results in Table 5 show that using the three criteria increases the average document separation by about 10% (relative), when compared to the Boolean weighting scheme. This result indicates that the document space is sparser when using the proposed importance function, which guaranties a better separation of the documents.

Weighting Scheme	*density*	Document separation increase
Boolean	0.249724	–
Importance	0.224163	10.23%

Table 5: Comparison of the document space density for a Boolean weighting scheme and the proposed weighting scheme.

[5] In our experiments, image collections have been annotated manually in order to ensure a high annotation quality, so that the experiments can evaluate the quality of the model without being biased by errors of an automatic annotation process.

5.3 Retrieval results

This experiment is aimed at comparing the retrieval accuracy of our model with the set of 20 queries. The model is implemented in a vector space system [32] in which documents are represented with vectors in a label space, where each coordinate denotes the importance of the corresponding label. We compare 3 weighting settings for the documents, namely a Boolean weighting scheme, our importance model (see the equation 1), and the importance model combined with the standard *idf* component. We also include the retrieval results of a standard color histogram search system. The results are given in Figure 7.

The result of the color histogram matching is very high for a low recall values because the first result is the query image itself, and then the following results are often some photographs of the same scene, taken from a slightly different angle. Then the precision drops drastically and is very low from about 20% of recall. This expectable result can be explained by the fact that visually similar images (close color histograms) are not necessarily semantically similar. The color histogram system can only retrieve results that are visually similar, even though their semantic content might be very different from the query.

When comparing the 3 weighting settings for the vector space system in Figure 7, we notice that they yield quite similar performances, with mean average precisions of 0.7094, 0.6893, and 0.7110 for *Boolean*, *Importance*, and *Importance* $\cdot$ *idf* settings respectively. By observing the detailed results for each query, we also notice that the difference between the settings are not statistically significant, under Wilcoxon two-sample test. This can be explained by the fact that, since the precision is very high for the 3 systems, the results contain a larger part of relevant documents, and therefore the differences consist mainly of **re-rankings among relevant documents**, which can neither be captured nor differentiated with a recall-precision plotting (see the discussion and example in the section 5.4).

In order to highlight this point, we have included in Figure 7 the expectable result of a text query-based search, where queries consist of single keywords corresponding to the label of the target objects. Since the images have been carefully indexed manually, the result for this search contains *all relevant images*, and *only relevant images*, yielding a precision of 1 for all recall values. Moreover, this result remains the same whatever weighting scheme is used. The reason for this is that the standard recall-precision measures evaluate systems based on a ground truth about the *binary relevance of documents*. For this reason, it is necessary to use another metric to evaluate the quality of the ranking **among relevant images**, that would take into account not only the *binary relevance of documents*, but also their rank in the ground truth. The next subsection presents such a metric, with the ranking results of our system.

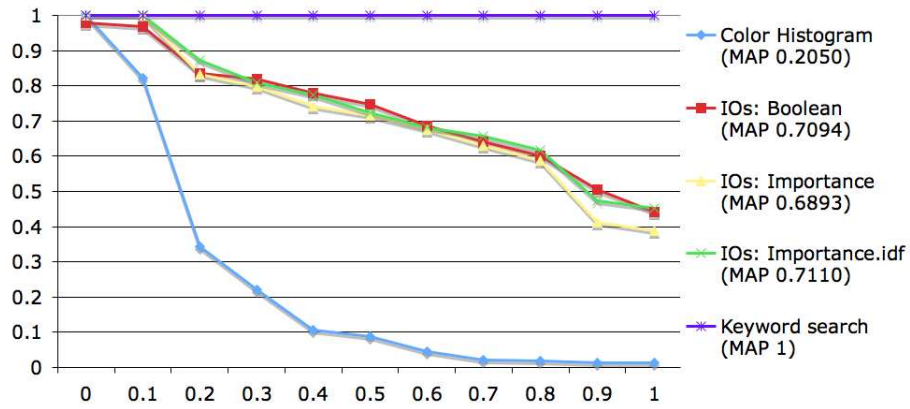


Figure 7: Recall-precision graph of the QBE experiment for the global color histogram matching, the IO matching with a Boolean weighting scheme, and the IO matching with the proposed weighting scheme – with and without the *idf* component. The graph also displays the result of a text-based keyword search.

5.4 Ranking comparison

The objective of this experiment is to evaluate precisely the ranking quality of the proposed weighting model with regard to users' aboutness assessments. For this purpose, we compare our system ranking of the images – in decreasing order of importance – to an assessors' ranking – in decreasing order of aboutness – with the same set of 20 queries in a text format, consisting of one single keyword corresponding to the label of the target object (similarly to the text query-based search reported in the previous subsection). The system ranking for each query is defined as the relevant images sorted in decreasing order of importance values for the keyword. The ideal user ranking has been generated by a group of 4 assessors[6] – 2 women and 2 men aged from 23 to 25 who have a good knowledge of the collection – and who have carefully ranked relevant images for each query (on average, 6 relevant images have been ranked for each query). An average ranking is generated by merging the individual rankings of the 4 assessors (as shown in [24]), which is used as a ground truth to which the system ranking is compared. The comparison to the ideal ranking will indicates how close to the users' perception of aboutness our weighting model is. This comparison is based on the following divergence function (inspired from [24]) that gives low divergence values to similar rankings, while penalizing more system ranking errors at the top ranked images:

$$div(U,S) = \frac{1}{mdv_n} \sum_{i=1}^{n} \frac{rk(U,i) - rk(S,i)}{rk(U,i)}$$

where U and S are the user ranking and the system ranking of n images respectively, $rk(U,i)$ and $rk(S,i)$ are the ranks of image i in the user ranking and in the system ranking respectively; mdv_n is the maximum divergence value for n items[7]. This divergence function gives the value 0 when $U = S$, and it gives the value 1 when U and S are in reverse order. For instance, the ranks for "Tree" are u=[5,8,1,4,6,3,2,7] and s=[5,8,3,2,4,6,1,7], meaning that the image 1 has been ranked at the 3^{rd} position by the assessors, while it is ranked at the 7^{th} position by the system. The divergence value for u and s is:

$$div(u,s) = \frac{1}{76.15} \cdot (\frac{16}{7} + \frac{9}{4} + \frac{9}{3} + \frac{1}{5} + \frac{0}{1} + \frac{1}{6} + \frac{0}{8} + \frac{0}{2})$$

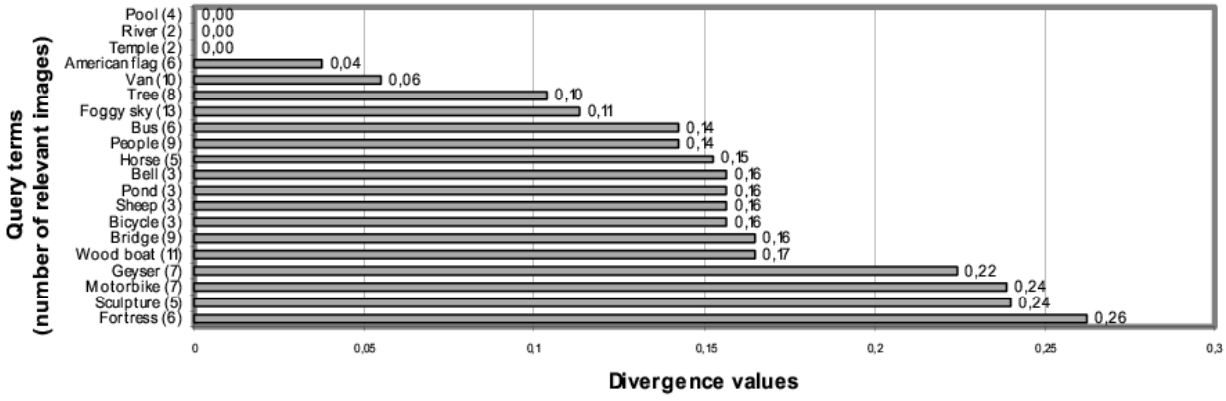


Figure 8: Divergence values for all query terms.

This value corresponds to a ratio of the ranking quality of a system over the worst ranking, which is the reverse order as compared to the ideal ranking. Note that a Boolean system not implementing any weighting scheme is likely to provide an arbitrarily random ranking, possibly related to the order in which relevant images are found by the algorithm in a list, which order may depend on the image file names in the system. Hence, this experiment provides a comparison between our system and a Boolean system not implementing any weighting scheme, that would return a reverse order compared to the ideal ranking.

[6] These 4 assessors are different from the ones who have participated in the first experiment.

[7] The maximum divergence value is reached when the two rankings are in a reverse order.

We use this divergence function in addition to the classical recall-precision measures because we are specifically interested in the order of relevant images, rather than in the usual binary relevance of images. Indeed, as discussed earlier recall-precision measures do not make it possible to differentiate two systems retrieving all relevant documents and only relevant documents, **in different orders**. For instance, if the result of one system (that is to say the sorted list of images) is s=[5,8,3,2,4,6,1,7], and the result of another system is s'=[5,8,1,4,6,3,2,7] (which is actually the same as u, the ideal ranking provided by the assessors), then the precision values for both systems are 1 for all possible recalls (see the keyword search result in Figure 7), since both systems retrieve *all relevant images, and only relevant images*. This example shows that in this specific experiment, the recall-precision measure is not able to point out that the ranking s' is better than the ranking s, and that, therefore, the second system performs better than the first one.

Note that the size of the collection and the number of relevant images per query are purposely kept small in our experiments so that images can be precisely annotated manually, and so that assessors can accurately rank images according to their relevance to queries. The system implements both each criterion individually and their combination. The divergence value (DV) averaged over the 20 queries is 0.24 when considering the surface only, 0.31 for the position, and 0.63 for the homogeneity. The latter DV is high because relevant images are sorted according to their homogeneity values, which do not take into account the query term. When combined together, the 3 criteria perform better that separately, as shown in Figure 8: DVs range from 0 (for *"Pool"*, *"River"* and *"Temple"*) to 0.26 (for *"Fortress"*). In Figure 8, the number of relevant images for each query is shown in parentheses. The results we obtained depend neither on the number of relevant images for the queries, nor on the type of objects considered (natural vs. non-natural, generic vs. specific or living vs. non-living). The high DVs (greater than 0.20) correspond to the queries *"Fortress"*, *"Geyser"*, *"Motorbike"* and *"Sculpture"*. For these query results, some of the relevant images are very similar with respect to their visual configuration, despite variations at an aesthetic level. Hence, the system is unable to discriminate between these images since the importance values, based on the visual configuration of images, are close to one another. However the average DV is 0.13, which means that our system ranking is very close to the users' perception of aboutness. These results lead us to conclude that our weighting model is adequate according to our second study. This experiment provides a validation to both our criteria modeling and their combination.

6 Conclusion

In this paper we have introduced a new framework of multimedia document description and matching dedicated to images and videos. This framework includes the definition of media objects and more specifically image objects, and the design of a model of importance for image objects. This model of importance is a contribution to image and video indexing and retrieval, that was viewed in this paper as a fundamental step for testing an image counterpart of the well-known *tf* paradigm in textual indexing. The model is fundamentally based on four perception criteria related to two-dimensional geometry of scenes, namely surface, position, fragmentation and homogeneity. These four criteria have been thoroughly confronted to actual human perception of images, and to aboutness assessments through an experiment involving real users. Three of these criteria have been experimentally validated in this process, formally modeled and further combined within a single weighting model for image objects. Finally, the effectiveness of this weighting scheme has been successfully compared to human-based aboutness assessments for images given symbolic definitions of topics. Though we understand that several other perception criteria could be also considered as a basis for such a model, we consider that we have already got a good basis for designing relevant similarity matching for images and videos, and made a step towards human-centered similarity matching for multimedia documents.

Future works are then mainly aimed at extending the weighting model to other perception criteria – like visibility of image objects related to contrast or luminosity – as well as exploring the impacts of motion (for video documents) and the context, in just the same experimental and formal way used for the four criteria described in this paper.

References

[1] S. Ayache, G. Quénot, and J. Gensel. Classifier fusion for svm-based multimedia semantic indexing. In *European Conference of Information Retrieval (ECIR'2007)*, pages 494–504, 2007.

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley, 1999.

[3] G. Baillargeon. *Probabilités, statistique et techniques de régression*. SMG, Québec, 1998.

[4] M. Bastan and P. Duygulu. Recognizing objects and scenes in news videos. In *Proceedings of CIVR'2006*, Phoenix AZ, 2006.

[5] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, and J. Malik. Blobworld: A system for region-based image indexing and retrieval. In *International Conference on Visual Information and Information Systems*. Springer, 1999.

[6] E. Hoerster, R. Lienhart, and M. Slaney. Image retrieval on large-scale image databases. In *Proceedings of CIVR'2007*, pages 17–24, Amsterdam, The Netherlands, 2007.

[7] L. Hollink, A. Th. Screiber., B. J. Wielinga, and M. Worring. Classification of user image descriptions. *Intl. Journal of Human-Computer Studies*, 61:601–626, 2004.

[8] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[9] Q. Tian J.H. Lim and P. Mulhem. Home photo content modeling for personalized event-based retrieval. *IEEE Multimedia, Special Issue on Multimedia Content Modeling and Personalization*, 10(4):28–37, 2003.

[10] J.H. Lee. Analyses of multiple evidence combination. In *SIGIR'97*, pages 267–276, Philadelphia, USA, 1997.

[11] J.-H. Lim. Photograph retrieval and classification by visual keywords and thesaurus. *New Generation Computing*, 18:147–156, 2000.

[12] J.-H. Lim. Building visual vocabulary for image indexation and query formulation. *Pattern Analysis and Applications (Special Issue on Image Indexation)*, 4(2/3):125–139, 2001.

[13] K.-H. Lothar. *Empiristic Theory of Visual Gestalt Perception. Hierarchy and Interactions of Visual Functions*. Koeln: Enane, 2001.

[14] D. Lowe. Object recognition from local scale-invariant features. In *Proceedings of ICCV*, pages 1150–1157, 1999.

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

[16] Y. Lu and H. Guo. Background removal in image indexing and retrieval. In *International Conference on Image Analysis and Processing (ICIAP 1999)*, pages 933–938, Venice, Italy, 1999.

[17] P. Martin and P. Bateson. *Measuring Behaviour: An Introductory Guide*. Cambridge University Press, 1986.

[18] J. Martinet, Y. Chiaramella, and P. Mulhem. A model for weighting image objects in home photographs. In *ACM-CIKM'2005*, pages 760–767, Bremen, Germany, 2005.

[19] J. Martinet, Y. Chiaramella, P. Mulhem, and I. Ounis. Photograph indexing and retrieval using star-graphs. In *Proceedings of CBMI'03 - Third International Workshop on Content-Based Multimedia Indexing*, pages 335–341, Rennes, 2003.

[20] J. Martinet and S. Satoh. Using visual-textual mutual information for inter-modal document indexing. In *ECIR'07*, Rome, Italy, 2007.

[21] M. Mitra, A. Singhal, and C. Buckley. Improving automatic query expansion. In *Research and Development in Information Retrieval*, pages 206–214, 1998.

[22] P. Mulhem, J.H. Lim, W.K. Leow, and M. Kankanhalli. Advances in digital home photo albums. In S. Deb (Ed.), editor, *In Multimedia Systems and Content-Based Image Retrieval*. Idea Group Publishing, 2003.

[23] W. Osberger and A. J. Maeder. Automatic identification of perceptually important regions in an image using a model of the human visual system. In *ICPR*, Brisbane, Australia, 1998.

[24] I. Ounis and M. Pasca. Finding the best parameters for image ranking: a user-oriented approach. In *Proceedings of The IEEE Knowledge and Data Engineering Exchange Conference (KDEX'98), Taipei, Taiwan*, pages 50–59, 1998.

[25] J. Preece. *Human-Computer Interaction*. Addison-Wesley, 1994.

[26] Y. Qiu and H.-P. Frei. Concept-based query expansion. In *SIGIR'93*, pages 160–169, Pittsburgh, USA, 1993.

[27] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, and T. Tuytelaars. A thousand words in a scene. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1575–1589, 2007.

[28] P. Quelhas and J.-M. Odobez. Natural scene image modeling using color and texture visterms. In *CIVR'2006*, Phoenix AZ, 2006.

[29] K. Rodden. How do people organise their photographs? In *Proceedings of 25th BCS-IRSG Colloquium on IR*, 1999.

[30] K. Rodden and K.R. Wood. How do people manage their digital photographs? In *ACM Conference on Human Factors in Computing Systems - CHI'03*, pages 409–416, Florida, USA, 2003.

[31] A.S. Rojet and E.L. Schwartz. Design considerations for a space-variant visual sensor with complex-logarithmic geometry. *10th International Conference on Pattern Recognition*, 2:278–285, 1990.

[32] G. Salton. *The SMART Retrieval System*. Prentice Hall, 1971.

[33] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. In *Information Processing and Management*, pages 513–523, 1988.

[34] G. Salton and M. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[35] G. Salton, A. Wong, and C. Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

[36] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[37] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[38] J. Sivic and A. Zisserman. Video google: a text retrieval approach to object matching in videos. In *Ninth IEEE International Conference on Computer Vision*, pages 1470–1477, vol.2, 2003.

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

[40] C.G.M. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35, 2005.

[41] F. Stentiford. An attention based similarity measure with application to content based information retrieval, 2003.

[42] B.M. ter Haar Romeny. *Front-End Vision and Multi-Scale Image Analysis*. Kluwer Academic Publishers, 2003.

[43] C.J. van Rijsbergen. *Information Retrieval, 2nd edition*. Butterworths, London, 1979.

[44] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

[45] J. Vogel and B. Schiele. Semantic modeling of natural scenes for content-based image retrieval. *Int. J. Comput. Vision*, 72(2):133–157, 2007.

[46] J.Z. Wang and Y. Du. RF x IPF: A weighting scheme for multimedia information retrieval. In *ICIAP*, pages 380–385, 2001.

[47] J.Z. Wang, J. Li, and G. Wiederhold. SIMPLIcity: Semantics-sensitive Integrated Matching for picture LIbraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947–963, 2001.

[48] I. Yahiaoui, B. Merialdo, and B. Huet. Comparison of multi-episode video summarisation algorithms. *EURASIP Journal on Applied Signal Processing*, 1:48–55, 2003.

[49] Q.-F. Zheng, W.-Q. Wang, and W. Gao. Effective and efficient object-based image retrieval using visual phrases. In *ACM Multimedia'2006*, Santa Barbara, California, 2006.