On the consistency and features of image similarity

Pierre Tirilly
School of Information Studies
University of Wisconsin-Milwaukee
Milwaukee, WI, USA
tirilly@uwm.edu

Chunsheng Huang
School of Information Studies
University of Wisconsin-Milwaukee
Milwaukee, WI, USA
huang22@uwm.edu

Wooseob Jeong
School of Information Studies
University of Wisconsin-Milwaukee
Milwaukee, WI, USA
wjj8612@uwm.edu

Xiangming Mu
School of Information Studies
University of Wisconsin-Milwaukee
Milwaukee, WI, USA
mux@uwm.edu

Iris Xie
School of Information Studies
University of Wisconsin-Milwaukee
Milwaukee, WI, USA
hiris@uwm.edu

Jin Zhang
School of Information Studies
University of Wisconsin-Milwaukee
Milwaukee, WI, USA
jzhang@uwm.edu

ABSTRACT

Image indexing and retrieval systems mostly rely on the computation of similarity measures between images. This notion is ill-defined, generally based on simplistic assumptions that do not fit the actual context of use of image retrieval systems. This paper addresses two fundamental issues related to image similarity: checking whether the degree of similarity between two images is perceived consistently by different users and establishing the elements of the images on which users base their similarity judgment. A study is set up, in which human subjects have been asked to assess the degree of the pairwise similarity of images and describe the features on which they base their judgments. The quantitative analysis of the similarity scores reported by the subjects shows that users reach a certain consensus on similarity assessment. From the qualitative analysis of the transcripts of the records of the experiments, a list of the features used by the subjects to assess image similarity is built. From this, a new model of image description emerges. As compared to existing models, it is more realistic, free of preconceptions and more suited to the task of similarity computation. These results are discussed from the perspectives of psychology and computer science.

Categories and Subject Descriptors

H.1.2 [**Models and principles**]: User/Machine systems—*Human information processing*; H.2.8 [**Database management**]: Database applications—*Image databases*

General Terms

Human Factors, Experimentation

IIiX 2012, Nijmegen, The Netherlands

Keywords

Image retrieval, Image similarity, Image description, Human perception

1. INTRODUCTION

Images are known to be a major challenge for information retrieval systems. Although tremendous progress has been made in recent years, as acknowledged by the development of general public applications such as Google Goggles[1] or Tineye[2], state-of-the-art systems are limited to very specific retrieval tasks such as near-duplicate detection and fail at providing effective search and browsing capabilities for general purpose collections, such as the ones provided on Flickr[3]. Such systems aim at providing *relevant* images to their users, but are actually mostly relying on the notion of *image similarity*. Indeed, measures of image similarity are used to retrieve images with respect to an image query [11], organize search results [19, 14] and automatically annotate images [15, 13]. However, this is a ill-defined notion. In image retrieval experiments, similarity is defined according to the image collection used to evaluate the image descriptors and the similarity measures. In this context, two definitions coexist:

- two images are similar if they are "images of the same object or scene viewed under different imaging conditions" [11];
- two images are similar if they depict objects or scenes from the same category [5, 16].

These definitions correspond to two of the three search behaviors of users observed by Cox *et al.* [3], respectively target search and category search[4]. However, other studies show that these categories are actually broader, as stated by Smeulders *et al.* [25] in their review. Another definition of similarity is the system similarity, *i.e.* the similarity as it is defined by the descriptors and similarity measures of

[1] www.google.com/mobile/goggles
[2] www.tineye.com
[3] www.flickr.com
[4] The third task is "search by association", *i.e.* image browsing.

the system. This is the definition on which annotation systems or browsing systems are based. Most datasets used to evaluate systems, hence the datasets from which similarity is defined, have known issues that make them poorly realistic [16, 18]. As a consequence, the notion of similarity as it is usually considered in the literature is defined artificially, out of the actual context of use of image retrieval systems. Authors like Cox *et al.* [3] and Jaimes *et al.* [10] suggested that more advanced models of image similarity validated on empirical data would benefit image retrieval systems.

This paper addresses the problem of defining the notion of image similarity *as perceived by the users*. More precisely, it aims at defining on which features people base their judgment of similarity. Previous work on image models [7, 8, 10] only addressed this issue in the context of image description, in which users tend to mention a very limited set of image features, as pointed out by Jörgensen *et al.* [12]. Before defining such features, though, it is necessary to check if the notion of similarity is common to the different subjects of the study. Indeed, establishing which features are useful to assess image similarity is pointless if the notion of similarity is too subjective, as this subjectivity will prevent the definition of any image indexing framework that would be both general and effective. As a result, the two following questions are successively addressed here:

1. Is the degree of similarity between two images perceived consistently by different users?
2. On which features of the images do users base their similarity judgments and how are these features related?

To answer these, an experimental study has be designed, during which human subjects are asked to rate the degree of similarity of pairs of images and provide the reasons that justify their ratings. The images used in this study have been selected to reproduce as faithfully as possible the conditions in which similarity would have to be assessed in a real-world image retrieval scenario. Although some previous studies also addressed this issue of pairwise similarity rating by humans [3, 6, 20, 23, 17], none of them addressed the problems of checking the consistency of human judgments and providing qualitative explanations of the ratings.

The paper is organized as follows. First, the experimental design of the study is presented (Section 2). This is followed by the analysis of the data, organized in two parts, corresponding to the two research questions addressed: the consistency of the degree of image similarity (Section 3), and the features used to assess consistency (Section 4). Finally, these results are discussed (Section 5) and the related literature is presented (Sections 5.2 and 5.3).

2. EXPERIMENTAL DESIGN

This section describes the design of the experiments, the data used, the recruitment of the subjects and the data gathered at the end of the experiments.

2.1 Experimental interface and scenario

Figure 1 shows a screen shot of the experimental interface that was developed. Based on this interface, the experiments follow this scenario:

1. The participant is asked to provide their informed consent by signing a consent form.

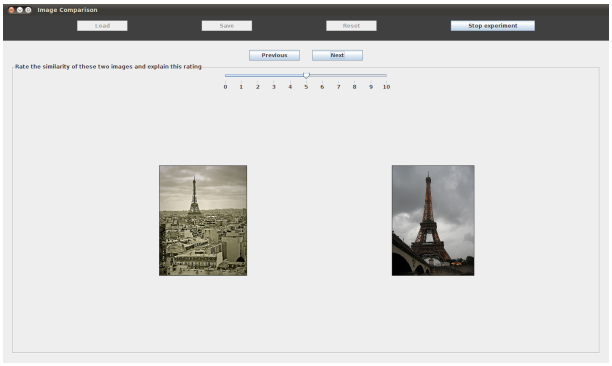

Figure 1: Screen shot of the experimental system used.

2. The investigator explains the task to the participant.
3. The participant is shown the first pair of images. He/she chooses a score between 0 and 10 reflecting his/her perception of the similarity between the images, then explains his/her decision following a think-aloud protocol. The next pair of images is then presented, and step 3 is repeated until no new pair of images is available. Figure 2 shows an example of the data obtained from one subject during this step.
4. The participants answer a post-study questionnaire to provide demographic information and a feedback about his/her experience during the study.

Precautions were taken to limit any bias during the experiments. No indication was given to participants regarding the kind of features that would be expected to explain their similarity judgments, so that they would be free to describe only what they have in mind. Presentation bias is avoided by displaying image pairs in a random order, and positioning images within a pair on the left or the right side of the screen at random.

2.2 Experimental data

The subjects are shown 40 image pairs during an experimental session. The images come from two sources: the MIRFlickr dataset [9], which is a dataset of 10,200 images downloaded from the image hosting website Flickr, and the LabelMe image dataset [21], a dataset developed by MIT. These datasets were chosen for three reasons:

- they are reference evaluation datasets used in multimedia information retrieval and computer vision;
- they contain images provided by professional and amateur photographers and designers, *i.e.* they correspond to real-world images like those that can be found in general image collections available on the web or in personal image collections;
- they come with user-generated annotations.

The annotations are used as a basis to generate image pairs for the experiments: images pairs are generated at random using images that have at least one annotation in common. Among those pairs, 40 have been manually selected for the experiments. This pair generation process has been used for following reasons:

Similarity score: 7

Both of horses. There's three horses in one. An-... One horse in the other. Looks like different breeds of horses. Different places. One's more of an open field, looks like. The other one looks... well, can't really tell, it's closeup. The one on the right seems like it's focused on the horse, the background's a bit blurred, whereas the other one's more of a general picture, there's no focus. I mean, besides the horses, but it's not like it's blurring the background.

Figure 2: Example of similarity assessment of two images by a subject.

- the image pairs have a relative thematic consistency. This prevents us from presenting to the subjects only completely unrelated images, which would not provide any significant data;

- the pairs correspond to images that would be presented together to a user submitting a query to a text-based image search engine.

Figure 3 shows some samples of the image pairs used for the experiments. As one can see, the fact that the images within a pair share some annotations does not necessarily imply that the images have a high level of similarity, and therefore allows to produce an experimental dataset with many variations, while corresponding to the response of a realistic text-based image retrieval system.

2.3 Subjects

35 subjects were recruited to participate in the study. Recruitment was performed through posters, flyers, email and snowball sampling. To prevent any bias in the responses of the participants, they were required to be native English speakers (to ensure that they have a sufficient vocabulary to describe their thoughts precisely without having to test their fluency), familiar with the use of computers (weekly use at least) and have no vision impairment (after eventual correction, glasses or contact lenses). They were also required to be at least 18 years old. The compliance to these requirements was self-reported by the subjects. They received a financial compensation for their participation in the study.

The subject population is varied in age (from 18-21 to more that 59 years old), gender (11 men and 24 women) and level of education (from high school to Ph.D.). As subjects were required to be native English speakers, the population sample is culturally consistent.

2.4 Data gathered during the experiments

Two types of data are obtained at the end of the experimental session: the rating of the similarity degrees of the image pairs, and the recordings of the subject explanations, that are subsequently transcribed. Each of these is analyzed in the next two sections.

3. QUANTITATIVE MEASURE OF IMAGE SIMILARITY

This section describes the statistical analysis of the similarity scores provided by the users during the experiments. The objective of this first part is to answer our first question: is there a notion of similarity that is, on average, common to all subjects?

(a) (b) (c) (d)

Figure 3: Samples of images pairs used in the experiments. Pairs (a) and (b) are images from the LabelMe dataset, and (c) and (d) from the MIR-Flickr dataset.

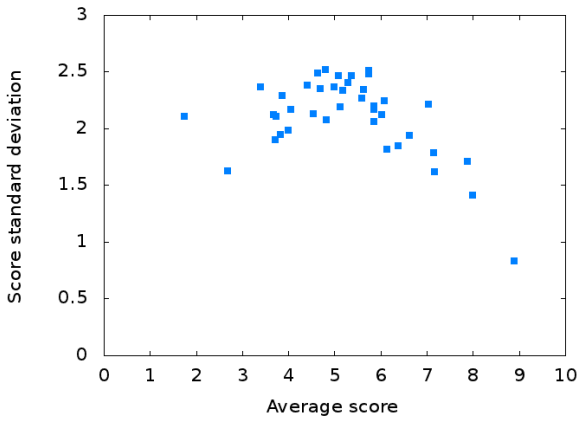

Figure 4: Standard deviation of the scores assigned to the pairs by the subjects with respect to their average value.

3.1 Analysis of raw similarity scores

Figure 4 shows the average (x-axis) and standard deviation (y-axis) of the scores assigned to each pair of images by the subjects of the experiments. Each point corresponds to an image pair. The points describe a bell-shaped curve: the standard deviation is lower for image pairs that have a low similarity score, and, more significantly, for pairs that have a high similarity score. Pairs that have an average similarity score also have a higher standard deviation. This shows that subjects tend to agree more on the similarity rating assigned to images recognized as being very similar or very dissimilar. This result suggests that an agreement between subjects exists only for these specific pairs of images. However, it should be noted that users can have subjective scoring scales. One major experimental factor explaining this variation in scoring scales is that the users tend to assess scores on the basis of the first pairs that they scored, which are used as a starting point to score the following pairs. This changes the average of the scores provided by the subjects. The range of scores assigned by the subjects also changes from one subject to another, for two reasons: some subjects tend not to use the whole scale available, and depending on the reference point chosen by the subject only a portion of the scoring scale might be available for further scorings (for instance, a subject that assigns a score of 7 to the first pair he/she sees has to rate the subsequent pairs perceived as more similar between 7 and 10). This effect can be avoided using a normalization of the scores.

3.2 Score normalization

To limit the subjectivity of the scoring scale used by each subject, the scores are normalized as follows:

$$\mathrm{NS}_{sp} = \frac{S_{sp} - \mu(S_s)}{\sigma(S_s)} \qquad (1)$$

where S_{sp} and NS_{sp} are respectively the score and normalized score assigned by subject s to image pair p, $\mu(S_s)$ is the average score assigned to all pairs by subject s and $\sigma(S_s)$ is the standard deviation of the scores assigned by subject s. Figure 5(a) shows the effect of this normalization to the set of scores assigned to all the image pairs by a subject.

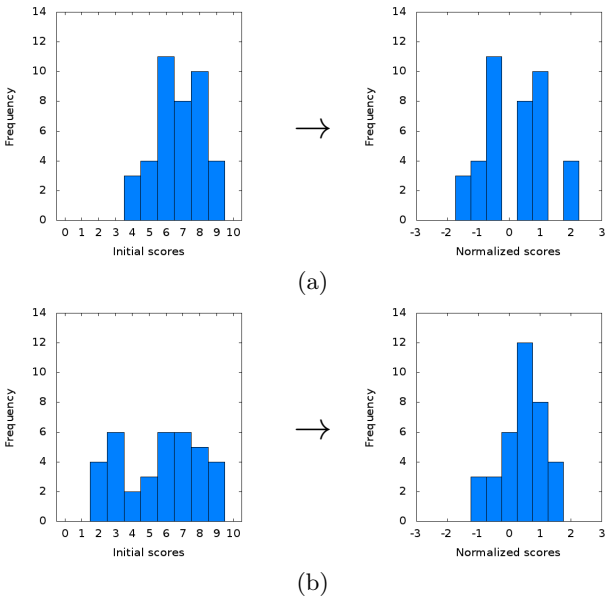

Figure 5: Histograms of the scores (a) assigned by a single user to all pairs (b) assigned by all users to a specific pair, before and after normalization of the scores.

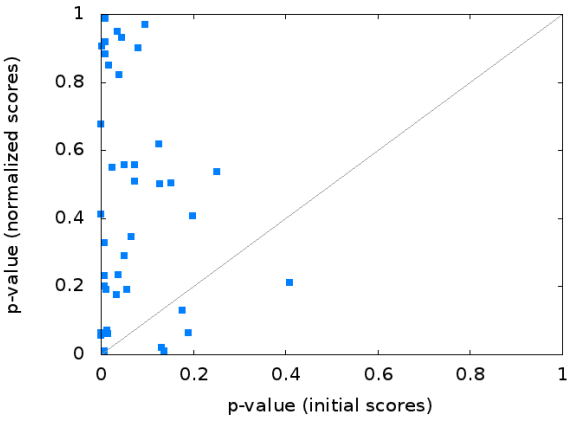

Figure 6: Comparison of the goodness-of-fit of the distributions of the image pair scores to a Gaussian before and after normalization.

The overall distribution is similar, *i.e.* the normalization did not change the relative scores, but has been translated according to the average and scaled according to the standard deviation. This avoids the two biases exposed in the previous paragraph: influence of the first scorings and subjective scoring scales. Rather than an absolute similarity rating, the normalized scores are relative scores that indicates if an image pair is more or less similar than the average (to this specific user). Figure 5(b) illustrates the effect of the normalization on the scores assigned by all users to a specific image pair. The initial scores are mostly uniformly distributed, whereas the normalized scores are distributed following a normal distribution. To confirm this effect, the

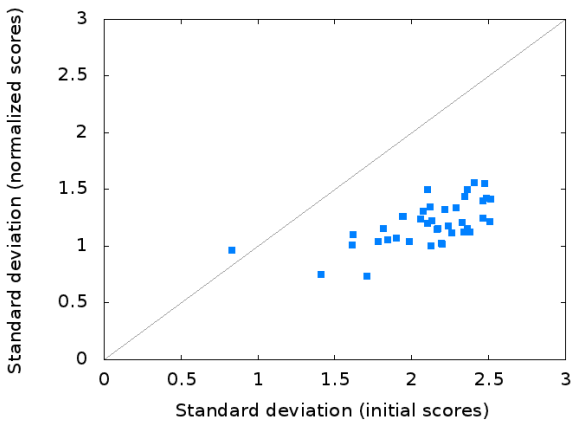

Figure 7: Comparison of the standard deviation of the similarity scores before and after normalization. Standard deviations of the normalized scores have been scaled by a factor of 5/3 to allow a fair comparison[6].

goodness-of-fit of the distribution of image pair scores to a Gaussian distribution has been computed using the Shapiro-Wilk test. The p-values of the test for initial scores (x-axis) and normalized scores (y-axis) are plotted in Figure 6. The p-value corresponds to the probability that the distribution is Gaussian. Pairs in the upper-left part of the plot have a better fit when normalized as compared to the initial scores. It shows that generally normalized scores follow a Gaussian distribution. As a consequence, image pairs tend to have a dominant similarity score (the average of their normalized scores), which can be considered as an "objective" similarity score. An analysis of the standard deviation of the distribution of pair scores (see Figure 7) shows that the normalization decreases the standard deviation of pair scores, and reveals that the agreement between the subjects is more important than it may seem at first.

In conclusion of this first part, this analysis of the similarity scores provided by the users for pairs of images highlight some properties of image similarity:

- subjects can agree on "absolute" similarity scores when the images are very similar and, to some extent, very dissimilar;
- once the effect of the subjectivity of the scores has been reduced, it appears that it is generally possible to identify a dominant similarity score for images.

4. THE ELEMENTS OF IMAGE SIMILARITY

This section describes the qualitative analysis of the transcripts of the recordings of the users. The objective is to address the second question stated in the introduction: on which features of the images do users base their similarity judgment and how are these features organized?

In the example of transcript obtained from a subject on the description of the similarity of two images presented in Figure 2, the subject states common points of the images (*e.g.* "Both of horses") and differences (*e.g.* "Looks like different breeds of horses. Different places."). These common points or differences corresponds to things that the subject sees (horse) or deduces from what he/she sees (breeds of horses, location). We will refer to these as image features in the remainder of the paper.

To answer the question addressed here, the image features mentioned in the transcripts are listed and organized into a model of image description. In the remainder of this section, this model is first presented, then compared to other models of image description that exist in the literature.

4.1 Image features

Table 1 presents the image features that were identified in the transcripts of the experiments. They are illustrated by one or more examples that occurred in the transcripts. To identify them, two of the authors of this paper went through the transcripts independently, listing the features they thought were relevant, then discussed their findings to come up with this final list of features.

The features listed here have the property to be objective features, *i.e.* features on which any two persons can agree (presence of objects, specific location...). Emotions are considered as objective, as Schmidt *et al.* [24] reported the consistency of the emotions felt by subjects submitted to individual images. From the transcripts that were analyzed only one notion was identified as being subjective: the notion of subject matter of the picture, *i.e.* stating what the image is about (and its natural complement, the notion of image background). For instance, in Figure 2, the subject clearly considers the horses as the subject of the picture, as opposed to the rest, which is considered as the background. Although there seemed to be a large agreement on this notion, it could happen that some subject had different opinions on what would be the subject matter of the picture.

4.2 Model of image description

From these features, an model of image description emerges (see Figure 8). First, the features can be organized into three levels of image description that clearly appear in the feature list (Table 1):

- the Object level, that relates to each individual entity present in the scene, physical or abstract;
- the Scene level, that relates to the whole set of objects that appear in the image and their shared properties (location, time...);
- the Image level, that relates to the image as a product.

Each of these levels contains some unique properties. The Object level contains all the properties related to individual entities that occur in the image. For instance, in the transcript excerpt "The bug is orange, it's furry but its antennas are much smaller, the wings are brown [...]", the color (orange, brown) and/or texture (furry) descriptions are clearly related to specific objects (bug, wings). These properties can be categorized into two categories, physical and semantic properties. Physical properties (color, texture, shape)

[6]Based on the fact that 99% of the values of a Gaussian distribution are within 3 standard deviations of the average, a scaling factor of 5/3 ensures that 99% of the normalized scores fall into a $[-5; 5]$ range equivalent in scale to the $[0; 10]$ range of the initial scores.

Image features	Examples of occurrence
Image Type	"Images of sailboats. One is a photograph and one is a drawing.", "The one on the right is just some type of a line drawing"
Image Focus	"The focus of this picture seems to be more of the scenery, although this one's directly on the bird.", "the focus of both [images] is the do not enter sign"
Image Point of View	"they are different angles from different perspectives looking at the same structure", "one's taken at just different angle and different distance away"
Image Lighting	"The lighting on this one is yellow, the one on this one is more... darker I guess.", "the lighting is brighter"
Image Contrast	"There is a lot more contrast in the left one I think, with the colors, and the one on the right is very congruent colors I think, it's kind of hard to tell.", "the background is darker so it kind of has more contrast"
Image Color	"[One image]'s black and white, one's in color.", "The picture on the right was also edited [...] all the colors were enhanced"
Image Quality	"the [image] on the left looks like it come from a magazine. The one on the right looks just like somebody [...] took a picture of their room.", "the [image] on the right strikes me as more of a professional photo, almost like from a catalog
Image File Quality	"It's also fairly pixellated, so it looks like it was a much smaller picture that was made larger"
Scene	"They're both street scenes", "one is a beach scene", "Looks like we have two bedroom spaces"
Scene Time	"Both night time", "the image on the right appears to be during the summertime"
Scene Location	"It's in Paris", "One of them looks like it's in the US, one of them looks like it's in another country"
Scene Event	"This one looks like a car show.", "Looks like Halloween."
Scene Purpose	"Appears to be a child's playroom. [...] The other appears to be a grownup's bedroom"
Scene Color	"The color schemes are in the earth tones versus.. water tones, blue tones", "The [bedroom] on the left has got a more blue-ish hue to it and the one on the right is very yellow."
Scene Composition	"[The images] both seem to be of just the pumpkin next to some shrubs", "There's three horses in one. An-... One horse in the other."
Scene Emotion	"They're sort of sense of foreboding", "They're very different, just the emotions, or feeling, that you get from both."
Object	"a baby", "a house", "a bike", "a woman", "horses"
Object Action	"They're both birds [...] One's in flight. One is perched on a stick.", "older man wearing black playing the guitar"
Object Purpose	"customize roaster specifically meant for the dragstrip", "advertising light up sign for [...] a barbecue"
Object Name	"Two photos of the Venus de Milo.", "The other one looks like the Winchester Mystery House."
Object Emotion	"The baby looks happy."
Object Nature	"the one on the left is dead [...] The one on the right is actually alive", "they're both of houses, but the one on the left is real. The one on the right is a Lego house."
Object Color	"the [moth] to the left is orange and yellow", "a white plate with a yellow apple on a brown placemat"
Object Texture	"Both have wooden frames, both have wooden furniture.", "The bug also has furry legs."
Object Shape	"the food is shaped into a person", "heart-shaped marking"
Object Composition	"You can see the eyes [of the bug], it has like leaf-like antennas.", "[the horse] has a dark brown mane"

Table 1: Examples of occurrences of the image features of the description model in the transcripts of the experiments.

correspond to low-level characteristics that do not require interpretation to be identified, whereas semantic properties correspond to higher level features such as action or purpose that require contextual information to be identified. This separation between the physical and semantic aspects of image interpretation is classically employed in image description [10, 8]. More interestingly, it corresponds to two approaches of image retrieval systems, namely content-based retrieval based on low-level features, and semantic retrieval based on keywords. At the Scene level, the properties can be organized in the same way as at the Object level, because the distinction between physical and semantic properties that apply to object properties also apply to scene properties, as a scene is basically a set of objects. However, individual properties may differ from one level to another, in the sense that some properties necessarily apply only to the whole scene, *i.e.* all the objects, for instance the location or the time of the scene. Finally, the Image level is related to any property of the image that is due to the choices of the image creator, for instance the point of view of the image, the focus, the type of image (photography, drawing) or the set of colors used (especially, black and white versus color images). These are all purely technical properties that are considered as a single category of image properties. Note that, although these features are not directly related to the objects that appear in the picture, they modify the way in which these objects are perceived by the observer.

In addition to having respective properties, the different levels are interrelated. By definition, the image contains a scene, which is what is represented in the image. The relation between scene and object exists through one property of the scene, scene composition. Indeed, when a subjects says "[The images] both seem to be of just the pumpkin next to some shrubs", it clearly describes the composition of the scene by stating the objects that the scene contains. Following the same idea, the object level is recursively connected to itself through the composition of the object, which contains itself sub-parts considered as individual objects (effect referred to as *Droste-effect* [8]).

This results in a fully interconnected set of features, organized into three levels, each level containing specific features.

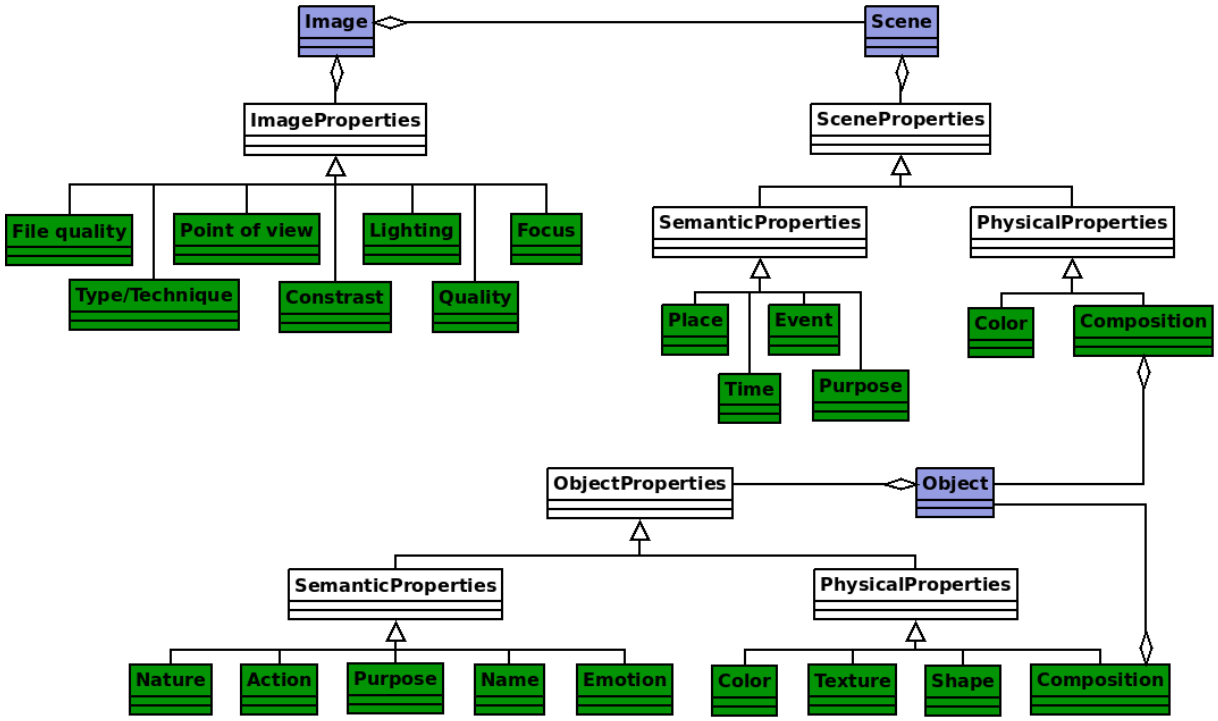

Figure 8: Model of image description. The diagram uses the UML formalism, with feature levels shown in blue and property features in green.

5. DISCUSSION

5.1 Limits of the model

Because the proposed model is built based on features obtained from experimental data, it is limited to the features that were actually relevant to the images used during the experiments. Although most features were sufficiently represented in the transcripts, and the image pairs were diverse, some features may be missing. More data would be necessary to validate the completeness of the model. It should be noted that it still contains more features than most models of image description [7, 10]. One specific kind of features that is missing is metadata, the data about the images (author, data of creation...). Metadata is absent from the transcripts as it was not presented to the subjects. However, the list of possible metadata is easy to obtain (for instance from [8]) and the model could be easily extended to include metadata as an additional type of properties at the Image level. Another limit is the fact that some features are closely related visually (*e.g.* image lighting and scene color), which makes them difficult to implement in an automatic system. However, this does not show that the model is not sound but that may have to be adapted according to the technical constraints or limitations of the system.

5.2 Differences with previous models

To our knowledge, three other models of image describing have been proposed previously [7, 10, 8]. The model proposed in this paper is different both in terms of methodology and organization. Methodologically, this model is purely empirical, based on actual statements from human subjects, whereas the existing models were built in a purely conceptual fashion, independently from any data, although they have been used *a posteriori* to analyze experimental data [7, 12, 8]. Thus, it corresponds to actual features and is not biased by pre-existent concepts with no empirical justification. Moreover, the model is based on an image comparison tasks, which is the core task performed by image retrieval systems. This model provides an accurate basis to define the image features that image retrieval systems should implement to reach their goal. It is also more complete as subjects comparing images will provide more details than subjects describing them. For instance, an image of a car in a street could be described simply using the words "car" and "street", whereas comparing two such images requires to making statements about their color, their model, their position... This lack of details was observed by Jörgensen *et al.* [12] in experiments on image description.

In terms of the resulting organization of the features, two major differences with previous models can be noted. First, this model associates objects or scenes with their specific properties (color, purpose...), whereas previous models separate them; especially, physical features are related to specific scenes or objects, whereas they are usually separated in other models, as authors generally organize their models into "perceptual" and "semantic" features by reference to psychological studies [10, 8]. Such a separation did not appear in the transcripts, where physical (or perceptual) properties are always related to a specific, well identified, object. This specific point is discussed in Section 5.5, and its consequences for image retrieval in Section 5.6. Moreover, the model proposed here does separate generic and

specific objects, scenes or concepts[7], unlike the models of Jaimes *et al.* [10] and Hollink *et al* [8]. As stated by Enser *et al.* [4], this distinction is not relevant as specificity is a gradual notion (for instance, *red car* is more specific than *car*, and *red corvette* more specific than *red car*). This notion of specificity has not been mentioned at all during the experiments. However, it is implicitly embedded in the model, as objects/scenes have properties, which accounts for their specificity or the specificity of their description. This offers a fine way to define the specificity of objects/scenes, made possible by the fact that objects/scenes are directly related to their physical or semantic properties. This does not provide information about the scale of specificity or how the different object/scene properties are located on this scale. This notion is out of the scope of the current study and would be unnecessary for a retrieval system that would be able to provide a fine calculation of image similarity, in which the notion of specificity is embedded.

5.3 Other related work

Some studies addressed the problem of visual similarity, either in the field of psychology or the field of computer science. In psychology, the objective of such studies is to explore how similarity is assessed at the lower level of the human vision system to understand how humans categorize images (*e.g.* [28, 1]). To do so, they base their experiments on simplistic images (single-line drawings) or specific images (*e.g.* faces), but do not work with images of complex scenes. As a consequence, they do not address similarity at a semantic level. As observed in the data presented here, and argued in Section 5.5, image similarity in the context of image retrieval should be seen at this semantic level, making psychological findings about low-level human vision not relevant in our context.

In computer science, a few authors have addressed the problem of the human assessment of similarity in the early years of image retrieval systems [3, 6, 20, 23, 26, 17]. In most cases, they did not try to check the consistency of this notion among subjects, nor try to explain which image features were relevant to describe image similarity. Rather, they used the data from their experiments to train machine learning algorithms that would try to reproduce the similarity assessments provided by the users. The fact that, as one can see in this work, the notion of similarity is complex and involve many semantic features that are not easily extracted from images even today, makes this learning problem difficult and most likely explains that this approach has been abandoned. The work by Squire and Pun [26] is likely the closest to ours: they check the agreement between human subjects on an image categorization task. The task is performed by subjects from the general public and experts in computer vision. They observe that only expert subjects agree on the categories to form, concluding that the perception of similarity is not naturally consistent but can be learned. The fact that they base their experiments on a classification task makes them more prone to disagreement between subjects, as the differences between the subjects can be due to many different images, which multiplies the reasons for these difference. Moreover, their conclusion about the agreement between experts must be considered carefully as it may depend on some strong preconceptions about image similarity that exist in the computer vision community (similarity generally considered as based on color, texture, shape, in the way that they try to implement it in their systems). Another related work is the one by Neumann and Gegenfurtner [17], who compare the similarity as perceived by users with similarity as computed by computers based on traditional low-level features (color and texture histograms). They conclude that the correlation exists, which would suggest that low-levels features are sufficient for indexing purposes. However, the fact that they use images from Corel collections, known for their positive bias towards image retrieval systems, makes their conclusions questionable. They also only present to their subjects images that were selected by the system, leaving out many images that were not retrieved but may have been more similar than the one they present.

Finally, this work could be related to experiments about relevance feedback, in which users of systems are asked to provide some judgments of relevance about the results provided by the system [29]. On the basis of this information, the system adapts its similarity measure to better fit the expectations of the user. In this case, as for similarity-based systems, no author, to our knowledge, provided any comprehensive analysis of the relevance information provided by the users. This information is used solely as input data for the system and the performance of the relevance feedback algorithm evaluated quantitatively. Moreover, as stated in Section 5.4, relevance is a very subjective notion, as opposed to similarity. As a consequence, it is likely that studying relevance feedback logs would provide information that would be different from the results presented here, and less consistent from one user to another.

5.4 Similarity versus relevance

As stated in the introduction, finding relevant images is the objective of image retrieval systems, whereas computing the similarity between images and queries is the process used to reach this goal. This paper addresses the problem of assessing similarity rather than relevance for two reasons:

- relevance is a complex notion that involves many parameters, among which many are subjective (previous knowledge, nature of the retrieval task...) [22]. As a consequence, two users providing the same query to a system might have very different expectations in terms of relevant documents, which makes it tricky to draw conclusions from the analysis of relevance-related data. By comparison, as confirmed by the results reported here, similarity seems to be more of an objective notion on which an agreement is possible. Indeed, when seeing a query and similar search results, two users may agree on some common characteristics or differences of the images, while at the same time one might consider a specific result as useful and the other not;

- as a consequence, trying to make image retrieval systems focus on relevance seems illusory, as the only input of the system is the query of the user, out of any context of use. Relevance feedback techniques might be used, but only within each search session of the user. Moreover, such efficient relevant feedback would also need a good similarity measure to be able to filter non-relevant results on the basis on the user judgments.

[7] An object, scene or concept is called specific if it can be uniquely identified by a named entity (*e.g.* The holy grail or World War II).

As compared to relevance, similarity offers a better basis for the development of systems that, instead of trying to provide the relevant documents directly, would provide images organized by similarity among which the user would navigate to reach his/her personal goal.

5.5 Psychological perspective

Research in psychology, and especially research about the human visual system (*e.g.* [28, 1]), offers a good basis for research about image retrieval and computer vision. As mentioned in Section 5.2, existing models of image description often refer to psychological studies, and especially make the difference between perceptual features (like color, texture or shape), that would be interpreted by the brain at early levels of the visual system, and semantic features, that would only be available at higher levels, after interpretation of the image [8, 10]. The model proposed here does not consider this separation as fundamental. Indeed, the experimental data shows that subjects assess similarity using semantic or perceptual features at the same level, in relation with specific objects or sets of objects of the image. This may indicate that the assessment of similarity for complex scenes is a high-level process only, which takes place after recognition of the content of the image. This is somewhat consistent with what Jörgensen *et al.* observed from users spontaneously describing pictures: they tend not to use perceptual features at all. This topic would require further analysis beyond the scope of the current paper.

5.6 Consequences for automatic image retrieval systems

The image model proposed here aims at being integrated to image retrieval systems. It shows which features should be indexed by the systems to provide a human-like similarity measure. Interestingly, current systems do not compute most of them, or use only a part of them, which may explain their limitations. Indeed, current systems generally fall into one of the following categories:

- content-based retrieval systems, that compute similarities based on low-level features only and ignore all semantic features [11, 25];
- annotation systems, which automatically associate key words to images, based on low-level features and machine learning [15]; promoters of such systems would consider that once the image is annotated, the retrieval problem is solved. The experiments presented here tend to show that, once the images are annotated, the retrieval problem just starts.

None of these systems try to combine low-level and high-level features, probably because some might consider that once the low-level features have been used to predict annotations, they might not be useful anymore. A few systems use both textual and visual information, fusing both information to get global descriptors, either by fusing retrieval scores or using one modality to filter the results obtain by the others [27, 2]. Such approaches correspond more to the findings of this study. Especially, the importance of combining semantic and physical features has been highlighted by the good performance of the system by Clinchant *et al.* [2]. However, all these systems rely only on global descriptions of the images, which prevents local associations between objects and their properties as described in users' similarity assessments.

Considering the analogy with the psychological remarks in the previous section, we could say that current systems only consider the lower levels of vision, from perception to the identification of object names[8]. As suggested above, the comparison of complex scenes might as well require higher level reasoning, which opens a whole new perspective on the development of image retrieval system.

6. CONCLUSION

Image similarity is at the center of image retrieval systems, but little is know about how humans assess the similarity in the case of complex images. This raises issues for the design and evaluation of real-world generic image retrieval systems. This paper addresses this issue by proposing a study where human subjects are asked to rate the similarity of pairs of images and describe the reasons for these ratings. The analysis of the experimental data results in two contributions:

- the analysis of the similarity scores show that a relative consensus exists between subjects about the rating of the similarity of image pairs and dominant values of the similarity ratings emerge once the influence of the subjective scales of users is removed;
- the analysis of the transcripts of the recordings allow us to list the features used by the subjects to assess the similarity. These features are organized into a model of image description that provides a better and more realistic basis to the development of image retrieval systems than previous models.

Future work includes a statistical analysis of the use of the image features by the subjects to establish which features are the most significant in the assessment of image similarity. It would also be interesting to gather more data of the same type, to confirm the present findings on a larger scale and develop an evaluation dataset for image retrieval that would be based on human judgments of similarity rather than artificial relevance criteria.

7. ACKNOWLEDGMENTS

This research was supported by the Research Group in Information Retrieval (rGIR) of the University of Wisconsin-Milwaukee.

8. REFERENCES

[1] I. Biederman. Recognition-by-components: A theory of human image understanding. *Psychological review*, 94(2):115–147, 1987.

[2] S. Clinchant, J. Ah-Pine, and G. Csurka. Semantic combination of textual and visual information in multimedia retrieval. In *Proceedings of the ACM International Conference on Multimedia Retrieval (ICMR)*, Trento, Italy, 2011.

[3] I. J. Cox, M. L. Miller, T. P. Minka, T. V. Papathomas, and P. N. Yianilos. The Bayesian image

[8] Which is acknowledged by the fact that they often refer to psychological studies and processes to justify their computational approaches.

retrieval system, PicHunter: theory, implementation and psychological experiments. *IEEE Transactions on Image Processing*, 9(1), January 2009.

[4] P. G. Enser, C. J. Sandom, J. S. Hare, and P. H. Lewis. Facing the reality of semantic image retrieval. *Journal of Documentation*, 63(4):465–481, 2007.

[5] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, April 2007.

[6] T. Frese, C. A. Bouman, and J. P. Allebach. A methodology for designing image similarity metrics based on human visual system models. In *Proceedings of SPIE, Human Vision and Electronic Imaging II*, pages 472–483, San Jose, CA, United States, February 1997.

[7] H. Greisdorf and B. O'Connor. Modelling what users see when they look at images: a cognitive viewpoint. *Journal of Documentation*, 58(1):6–29, 2002.

[8] L. Hollink, A. Schreiber, B. Wielinga, and M. Worring. Clasiffication of user image descriptions. *International Journal of Human-Computer Studies*, 61(5):601–626, 2004.

[9] M. J. Huiskes and M. S. Lew. The MIR Flickr retrieval evaluation. In *Proceedings of the ACM Conference on Multimedia Information Retrieval (MIR)*, pages 39–43, Vancouver, BC, Canada, October 2008.

[10] A. Jaimes, R. Jaimes, and S.-f. Chang. A conceptual framework for indexing visual information at multiple levels. In *in proceedings of SPIE, Internet Imaging*, pages 2–15, San Jose, CA, United States, January 2000.

[11] H. Jégou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proceedings of the European Conference on Computer Vision (ECCV)*, volume 1, pages 304–317, Marseille, France, October 2008.

[12] C. Jörgensen, A. Jaimes, A. B. Benitez, and S.-F. Chang. A conceptual framework and empirical research for classifying visual descriptors. *Journal of the American Association for Information Science and Technology*, 52(11):938–947, 2001.

[13] L. Kennedy, M. Slaney, and K. Weinberger. Reliable tags using image similarity: mining specificity and expertise from large-scale multimedia databases. In *Proceedings of the 1st workshop on Web-scale multimedia corpus (WSMC)*, pages 17–24, Beijing, China, 2009.

[14] H. Liu, X. Xie, X. Tang, Z.-W. Li, and W.-Y. Ma. Effective browsing of web image search results. In *Proceedings of the ACM International Workshop on Multimedia Information Retrieval (MIR)*, pages 84–90, New York, NY, USA, 2004.

[15] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 316–329, Marseille, France, May 2008.

[16] H. Müller, S. Marchand-Maillet, and T. Pun. The truth about corel-evaluation in image retrieval. In *Proceedings of the Conference on Image and Video Retrieval (CIVR)*, pages 38–49, London, UK, July 2002.

[17] D. Neumann and K. R. Gegenfurtner. Image retrieval and perceptual similarity. *ACM Transactions on applied perception*, 3(1):31–47, January 2006.

[18] J. Ponce, T. Berg, M. Everingham, D. Forsyth, H. M., S. Lazebnik, M. Marszalek, C. Schmid, B. Russell, A. Torralba, C. Williams, J. Zhang, and A. Zisserman. *Dataset Issues in Object Recognition*, pages 29–48. Springer-Verlag, 2006.

[19] K. Rodden, W. Basalaj, D. Sinclair, and K. Wood. Does organisation by similarity assist image browsing? In *Proceedings of the ACM SIGCHI conference*, Seattle, WA, United States, March-April 2001.

[20] B. E. Rogowitz, T. Frese, J. R. Smith, C. A. Bouman, and E. Kalin. Perceptual image similarity experiments. In *Proceedings of the SPIE, Human Vision and electronic Imaging III*, San Jose, CA, United States, January 1998.

[21] B. C. Russel, A. Torralba, K. P. Murphy, and W. T. Freeman. Labelme: a database and web-based annotation tool for image annotation. *International Journal of Computer Vision (IJCV)*, 77(1-3):157–173, May 2008.

[22] T. Saracevic. Relevance reconsidered. In *Proceedings of the second conference on conceptions of library and information science (COLIS)*, pages 201–218, Copenhagen, Danemark, October 1996.

[23] B. Scassellati, S. Alexopoulos, and M. Flickner. Retrieving images by 2d shape: a comparison of computation methods with human perceptual judgments. In *Proceedings if SPIE, Storage and Retrieval for Image and Video Databases*, pages 2–14, San Jose, CA, United States, February 1994.

[24] S. Schmidt and W. G. Stock. Collective indexing of emotions in images. A study in emotional information retrieval. *Journal of the American Society for Information Science and Technology*, 60(5):863–876, 2009.

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

[26] D. M. Squire and T. Pun. A comparison of human and machine assessments of image similarity for the organization of image databases. In *Proceedings of the Scandinavian conference on image analysis*, Lappeenranta, Finland, June 1997.

[27] S. Tollari and H. Glotin. Web image retrieval on ImagEVAL: Evidences on visualness and textualness concept dependency in fusion model. In *Proceedings of the Conference on Image and Video Retrieval (CIVR)*, pages 65–72, Amsterdam, The Netherlands, July 2007.

[28] A. Tversky. Features of similarity. *Psychological review*, 84(4):327–352, July 1977.

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