

CAPTURING IMAGE SEMANTICS WITH LOW-LEVEL DESCRIPTORS

Aleksandra Mojsilovic, Bernice Rogowitz
IBM TJ Watson Research Center, Hawthorne, NY 10532

Abstract: We conducted psychophysical experiments to gain insight into the semantic categories that guide the human perception of image similarity. We analyzed the perceptual data using multidimensional scaling (MDS) and hierarchical clustering (HC). Based on this analysis we established the most important semantic categories in the perception of image similarity. We then used these data to discover low-level features that best describe each category. Finally, we devised an image similarity metric that embodies our findings and models the behavior of subjects in categorizing images and measuring their similarity. Our results can be used for enhancement of current image/video retrieval methods, better organization of large image databases, development of more intuitive navigation schemes, browsing methods and user interfaces.

1. INTRODUCTION

High-level semantic concepts play a large role in the way we perceive images and measure their similarity. Unfortunately, these concepts are not directly related to image attributes. Although there are many sophisticated algorithms to describe color, shape and texture features, these algorithms do not adequately model image semantics and therefore have many limitations when dealing with broad content image databases. Yet, due to their computational efficiency the low-level visual attributes are widely used for image retrieval, leaving the user interface of content-based retrieval (CBR) system with a task of bridging the gap between the low-level nature of these primitives and the high-level semantics people use to judge image similarity. In this work we take another approach in overcoming this gap and pose the following question: “*Is it possible to find correlations between the high-level semantics and low-level descriptors and use them to capture the semantic meaning of an image?*”

1.1 Previous work and our objectives

As a starting point in our work we used the data from our earlier experiments [1], in which subjects judged the similarity of 97 photographic images. The experiments produced similarity matrices that were analyzed with the multidimensional scaling algorithms. The analysis discovered two important dimensions in human similarity perception: *natural vs. man-made*, and *humans vs. non-humans*. This study also revealed that images were primarily grouped according to broad semantic categories. The objective of our work was to continue from these findings, enhance our understanding of these perceptual categories and devise an adequate similarity model. We performed several additional subjective experiments to measure the similarity of 196 photographic images. To analyze the experimental data we used MDS and HC. These experiments helped us discover the most important

categories in human similarity perception, their semantic relevance, organization and relationships. We then described the obtained semantic categories in terms of image processing operations and calculable image features. For each category we established a feature combination that captures its semantics and discriminates it from the other categories. Finally, we devised a similarity metric that embodies our findings, categorizes and retrieves images based on their semantics.

2. EXPERIMENTAL SETUP AND DATA ANALYSIS

2.1 Selection of stimuli and subjects

We used 196 digitized photographic images divided into two sets. The first set (*Set 1*) contained 97 images from the “PhotoDisk” collection. These images were used in the previous study [1] and the similarity data from that study was available for further analysis. The second set (*Set 2*) contained 99 images selected according to the following criteria. As in the previous experiment, we included a wide range of topics: people, nature, buildings, textures, objects, indoor scenes, animals, etc. For each topic we explicitly selected wide angle, normal and close-up images, in both landscape and portrait modes. We selected images that complemented the first set and also included some additional topics. Finally, we iterated our selection so that the set included a wide distribution of brightness levels and colors.

Seventeen subjects participated in the study. Their ages ranged from 24 to 65. All the subjects had full color perception. The subjects were not familiar with the images.

2.2 Selection of preliminary semantic categories

To determine the set of preliminary semantic categories we used the subjective data from [1] and performed the HCA. We then split the similarity data in several ways, eliminated some of the stimuli from the data matrix and reapplied the HCA for the remaining stimuli. The clusters that remained stable for various solutions were included into a set of preliminary categories (*PC*) to be used in our study. Images that did not cluster consistently in the different HCA’s were treated as separate clusters.

2.3 Experiment 1

For the first experiment we printed thumbnails of the 97 images from Set 1 and organized them on a tabletop by category, with a clear space between the categories. We also printed thumbnails of the images from Set 2. Twelve subjects participated in this experiment. They were asked to assign each image from the new set into one of the existing categories according to their perceived similarity.

Subjects were also asked to organize images in such a fashion that the most similar images were near each other. There were no instructions concerning the characteristics on which the similarity judgments were to be made. The presentation order was random and different for each subject. The subjects were not allowed to change the initial categories - these images were glued to the tabletop and could not be moved. However, subjects were allowed to do whatever they liked with the new images: to change their assignments during the experiment, keep images on the side and decide later, or start their own categories. At the end of experiment, the subjects were asked to explain some of their decisions. These explanations, as well as the relative placement within the categories, were used later as an aid in defining the final set of categories and their semantics.

2.4 Data analysis

The first step in the data analysis was to compute the similarity matrix Δ_{S_2} for images from Set 2. The matrix entry $\Delta_{S_2}(i, j)$ represents a number of times images i and j occurred in the same category. This matrix was used as an input to the MDS. The next step was to compute the similarity matrix $\Delta_{S_2,PC}$ for both the images from Set 2 and preliminary categories. The matrix entry $\Delta_{S_2,PC}(i, j)$ is computed in the following way:

$$\Delta_{S_2,IC}(i, j) = \begin{cases} \Delta', & i, j \in \text{Set 2} \\ \Delta'', & i \in \text{Set 2 and } j \in PC \\ d(i, j), & i, j \in PC \end{cases} \quad (1)$$

where: Δ' is number of times images i and j occurred in the same category, Δ'' is number of times image i occurred in the category j , and $d(i, j)$ is the Euclidean distance between the centers of the PC normalized to occupy the same range of values as similarity measures Δ' and Δ'' . This matrix was used as an input to the HCA to determine the final set of categories guiding the perception of image similarity. Finally, we transformed the similarity data into the confusion matrix CM , where entry $CM(c_i, c_j)$ represents an average number of images from the category c_i placed into the category c_j (and vice versa). We combined these values with the comments from our subjects and the HCA result to establish the connections between the categories.

2.5 Experiment 2

Having identified the final categories, we performed another experiment to investigate the correlation between the groupings and semantics. Images from each category were printed on one sheet of paper - the images were randomly organized on the top half of the sheet, while the bottom part was left blank. We organized these sheets into a *category notebook*. Ten subjects participated in this experiment: 5 of them took part in Experiment 1, while the other 5 subjects have not seen the images before and were not familiar with our objectives. Each subject was given the notebook and asked to name each category and write a brief description and main properties of that category. We used this experiment to establish if subjects perceive the

categories in a consistent manner. Furthermore, the experiment helped us understand the semantics and assign the name to each category. Finally, the written explanations were crucial in determining pictorial features that best capture the semantics of each category (Section 4).

3. EXPERIMENTAL RESULTS

3.1 Multidimensional scaling: Results

To test whether the dimensions discovered in [1] apply to our data we performed MDS to the similarity matrix Δ_{S_2} . The resulting 2D configuration is shown in Fig. 1. The stress was 0.21. As it can be seen, there is an excellent correspondence between our configuration and the one reported in [1] with the same interpretation for the dominant dimensions.

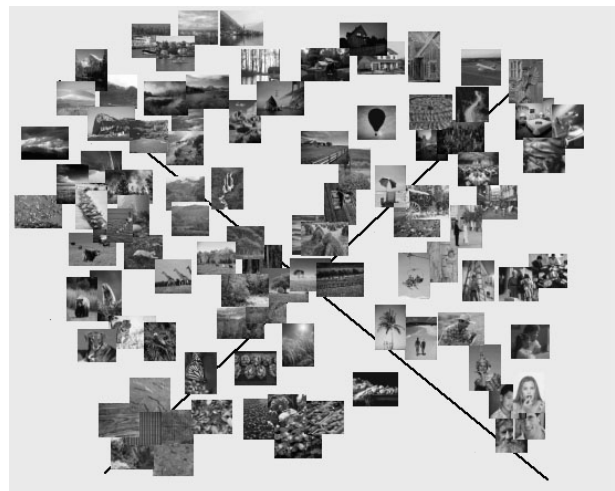


Fig. 1: 2D-MDS configuration, with the dimensions determined in [1].

3.2 Hierarchical clustering: Results

Human observers use many dimensions in evaluating image similarity. To capture the higher-dimensional nature of these judgments and to establish the most important semantic categories in similarity perception we performed HC. We named these categories using the written explanations of the subjects who participated in Experiment 2 and the verbal comments of the subjects who took part in the first experiment. These categories are: 1) Portraits, 2a) People outdoors, 2b) People indoors, 3) Outdoor scenes with people, 4) Crowds of people, 5) Cityscapes, 6) Outdoor architecture, 7) “Technoscenes”, 8) Objects indoors, 9) Objects outdoors, 10) Waterscapes with human influence, 11) Landscapes with human influence, 12) Waterscapes, 13) Landscapes with mountains, 14) Sky/Clouds, 15) Winter/snow, 16) Green landscapes, 17) Landscapes with fields and foliage, 18) Plants, flowers, fruits and vegetables, 19) Animals and 20) Textures, patterns and close-ups.

3.3 Qualitative findings

We were surprised to discover that some semantic descriptors were semantically more important than the others. For all subjects, “water”, “sky/clouds”, “snow” and “mountains” emerged as very important cues, often strongly related to each other, determining the organization and links between the groups. The same holds for images of people. We discovered that presence of people in the image provides much stronger cue than anything else, even in the case of a dominant natural scene, object or man-made structure. We also discovered that color composition and color features play an important role in comparing natural scenes. On the other hand, color was rarely used in judging the similarity of images with people, man-made objects or environments, where spatial organization, spatial frequency and shape features mainly influenced the similarity judgments. We also found that presence of strong colors (such as bright red, yellow, lime green, pink, pure white, etc.) can be used to indicate man-made objects in the picture, especially when combined with spatial and regional features and overall color composition. Image segmentation into regions of uniform color or texture yields opposite results for natural and man-made categories. Regions in “man-made” images have straight lines, straight boundaries, sharp edges, or geometric shapes; regions in “natural” images have rigid boundaries and more “random” edge distributions.

4. SEMANTICS AND PICTORIAL FEATURES

Our experiments confirmed that high-level semantic concepts are very important in judging image similarity. Although the semantic concepts are usually not directly related to the visual image attributes (color, texture, shape, etc.), these attributes frequently capture information about the semantic meaning [1]-[3]. Hence, we concentrated on various low-level image features and examined their correlation with the semantic categories. To do so, we used the written descriptions of the categories gathered in the second experiment and devised a list of verbal descriptors people found crucial in distinguishing the categories. We then translated these descriptors into calculable image-processing features. For example, the phrase “image consisting primarily of a human face, with little or no background scene” was used to describe the “Portraits” category. In “image-processing language” this corresponds to “dominant, large skin colored region”. In total, the feature list contained over 40 image-processing features, which we will call the *complete feature set* (CFS). Some of the features include: “number of regions after segmentation” (large, medium, small, one region), “energy” (high, medium, low frequencies), “central object” (yes, no), “color composition” (bright, dark, saturated, gray overtones, etc.), “blobs of bright color” (yes, no), “spatial distribution of dominant colors” (sparse, concentrated), “geometry” (yes, no), “number of edges” (large, medium, small, no edges), “straight lines” (occasional, defining an object, no straight lines), etc.

To find whether these features correlate with the semantics of each category, we used the Opal visualization tool [4]. We compared the experimental data with the

image-processing descriptors for a set of 100 images, and for each category determined a feature combination that discriminates that category from other images. For example, the following feature combination captures the semantics of the “Cityscapes” category:

Skin = no	Face = no
Silhouette = no	Nature = no
Energy = hi	Central object = no
Number of regions = large	Num. of edges = large
Region size = small/medium	Details = yes
Color = brown/gray.	

This process was repeated for all twenty categories. The detailed description of the feature sets can be found in [5].

We also discovered that not all the features within a certain category are equally important. For example, all “Cityscapes” have high spatial frequencies, many details, dominant brown/gray overtones and segmentation yields large number of small regions. These features are thus considered as *required features* for the “Cityscapes” category. In addition, many images from this category (but not all of them) have straight lines or regions with regular geometry, due to man-made objects in the scene. Similarly, although the dominant colors are on the brown/gray/dark side, many images have blobs of saturated colors, again due to man-made objects in the scene. We call these the *frequently occurring features* for the “Cityscapes” category.

5. CATEGORIZING AND RETRIEVING IMAGES

5.1 The metric for semantic categorization

Having discovered the similarity categories and their features our objective was to devise a similarity metric that embodies our perceptual findings and models the behavior of subjects in categorizing images. The metric is based on the following observations from our experiments: 1) There is a set of semantic categories people use in judging image similarity. 2) Each semantic category c_i , can be discriminated using the feature set $f(c_i)$:

$$f(c_i) = [RF_1(c_i) \dots RF_{M_i}(c_i) \quad FO_1(c_i) \dots FO_{N_i}(c_i)] \quad (2)$$

where: $\{RF_j(c_i) | j=1, \dots, M_i\}$ are M_i required features, and $\{FO_j(c_i) | j=1, \dots, N_i\}$ are N_i frequently occurring features for the category c_i .

To assign a semantic category to the input image x , we need a complete feature set $CFS(x)$. However, when comparing x to the semantic category c_i , we will use only a subset $f(x|c_i) \subset CFS(x)$ consisting of features that capture the semantics of category c_i :

$$f(x|c_i) = [RF_1(x|c_i) \dots RF_{M_i}(x|c_i) \quad FO_1(x|c_i) \dots FO_{N_i}(x|c_i)]. \quad (3)$$

The similarity between the image x and category c_i is computed via the following metric:

$$sim(x, ci) = \frac{1}{N_i} \prod_{j=1}^{M_i} \tau(RF_j(x | ci), RF_j(ci)) \cdot \sum_{j=1}^{N_i} \tau(FO_j(x | ci), FO_j(ci))$$

$$\text{where: } \tau(a, B) = \begin{cases} 1, & (\exists i) a = b_i \\ 0, & (\forall i) a \neq b_i \end{cases}, \text{ and } B = \{b_i\}_{i=1, \dots, I} \quad (4)$$

This metric represents a mathematical description of what we found so far: to classify an image into a semantic category, all the required and at least one of the frequently occurring features for that category have to be present. Since a required feature $RF(c_i)$ typically has more than one value (i.e. I possible values), the feature $RF(x|c_i)$ is compared to each possible value via (4).

We developed a Visual Basic application to test the metrics ability to accurately categorize images. For a selected image the application loads a precomputed CFS and determines the semantic category for that image. We tested the algorithm using 100 images from our experimental set and a manually filled feature table. The “perfect” feature table helped us gain better understanding of the metric by eliminating possible misclassifications due to feature extraction inaccuracies. The categorization was correct for 93/100 images from the set. Having validated the metric, the next step will be to implement automatic feature extraction algorithms.

5.2 Image retrieval based on semantic categorization

In addition to semantic categorization, the proposed metric can be used to measure similarity between two images, x and y as:

$$sim(x, y | ci) = \frac{1}{N_i} \prod_{j=1}^{M_i} \tau(RF_j(x | ci), RF_j(y | ci)) \cdot \sum_{j=1}^{N_i} \tau(FO_j(x | ci), FO_j(y | ci))$$

$$sim(x, y) = \max_i (sim(x, y | ci))$$

However, note that the similarity score is greater than zero only if both images belong to the same category. To allow comparison across all categories we propose a less strict metric. We first introduce the similarity between images x and y , assuming that both of them belong to the category c_i as:

$$sim(x, y | ci) = \frac{1}{2^{M_i+N_i}} \prod_{j=1}^{M_i} (1 + \tau(RF_j(x | ci), RF_j(y | ci))) \cdot \prod_{j=1}^{N_i} (1 + \tau(FO_j(x | ci), FO_j(y | ci)))$$

Assuming that $x \in c_i$ and $y \in c_j$, the overall similarity is defined as:

$$sim(x, y) = [sim(x, y | c_i) + sim(x, y | c_j)] / 2 \quad (5)$$

Fig. 2 shows one image retrieval example based on this metric.



Fig. 2: A retrieval examples based on the proposed metric.

5. CONCLUSIONS

The world of digital image libraries is growing rapidly, and more and more people will become interested in searching, navigating and browsing image collections. In order for this to be successful, we need to provide a user interface that is natural and intuitive. In this paper, we have presented the results of perceptual experiments aimed at understanding the way in which users judge the similarity of photographic images. The experiments demonstrate that image semantics play a large role in determining image similarity. We have used our perceptual findings to develop a semantically-based model of image similarity. To do so, we first identified broad semantic categories in the perceptual data, which we then modeled in terms of combinations of low-level image features. We have also developed a metric, based on these features, for categorizing and searching image databases.

Our results can be used for the enhancement of current image/video retrieval methods, for better organization of large image databases, and for the development of more intuitive navigation schemes, browsing methods and user interfaces.

6. REFERENCES

- [1] B. Rogowitz, T. Frese, J. Smith, C. A. Bouman, and E. Kalin, “Perceptual image similarity experiments”, *Proc. of SPIE*, 1997.
- [2] M. Turk and A. Pentland, “Eigenfaces for recognition”, *J. Cogn. Neurosci.*, vol. 3, pp. 71-86, 1991.
- [3] A. Mojsilovic, J. Kovacevic, J. Hu, R. J. Safranek, K. Ganapathy, "Matching and retrieval based on the vocabulary and grammar of color patterns", *IEEE Trans. on Image Proc.*, vol. 9, no. 1, pp. 38-54, January 2000.
- [4] D. Rabenhorst, *Opal: Users manual*, IBM Documentation.
- [5] A. Mojsilovic and B. Rogowitz, “A psychophysical approach to modeling image semantics”, *IST&SPIE Conference on Human Vision and Electronic*, San Jose 2001.