Assessing Agreement Between Human and Machine Clusterings of Image Databases [⋆]

David McG. Squire and Thierry Pun

Computer Science Department, University of Geneva, 24 rue Général-Dufour, 1211 Genève 4, Switzerland

Abstract

There is currently much interest in the organization and *content-based* querying image databases. The usual hypothesis is that image similarity can be characterized by low-level features, without further abstraction. This assumes that agreement between machine and human measures of similarity is sufficient for the database to be useful. To assess this assumption, we develop measures of the agreement between partitionings of an image set, showing that chance agreements *must* be considered. These measures are used to assess the agreement between human subjects and several machine clustering techniques on an image set. The results can be used to select and refine distance measures for querying and organizing image databases.

Key words: image similarity, perceptual distance, databases, query by content, clustering, reliability, expected agreement

1 Introduction

There has recently been a significant increase in the amount of research into methods for organizing and querying large image databases, containing many images of unconstrained real-world scenes. In particular, there has been great interest in developing means of querying such databases by *image content*, rather than by performing text-based matching of labels attached to images by a human expert [9,4,12,15,8,11,19]. The desire to use queries based on visual content stems from the belief that the proverb *"A picture is worth a thousand words"* encapsulates an important fact: simple text labels are inherently too

[⋆] This work is supported by the Swiss National Foundation for Scientific Research (grant no. 2100-045581.95).

terse to describe image contents adequately.[1] A measure of image similarity is required, both for evaluating the similarity of images in the database to a query image, and for organizing the database so that this search may be performed efficiently.

In this context, it is largely acknowledged that the object recognition problem for images of unconstrained environments remains unsolved [4]. This has lead many researchers to limit their work on content-based queries to databases of images from extremely restricted domains, such as industrial trademarks [9,8], or marine animals [14]. Even in these restricted domains, the approach has not been to attempt to extract a semantic description of the image. Rather, a variety of low-level image features has been employed, such as colour histograms [4], colour, segment and arc statistics [17,12], or shape measures for single object images [14,18]. In all these cases, the hypothesis has been that image similarity can be characterized by combinations of these low-level features, without it being necessary to move to a higher level of abstraction such as full-blown object recognition. It is assumed that there is sufficient agreement between machine and human measures of image similarity for the database to be universally useful. In this paper we will assess the veracity of this assumption.

In section 2 we discuss the need for a measure of image similarity in image database systems, propose an experiment for assessing human judgment of image similarity, and consider the ways in which such data could be used to rate and improve machine measures. In section 3 we show how a mathematical measure of the agreement between two partitionings of a set of images into unlabelled subsets can be defined, and develop statistics from this measure which indicate the degree to which the measured agreement is *better than that expected by chance*. The methodology of the experiment to assess the agreement of human subjects with each other and with a variety of machine clustering techniques is presented in section 4. The results of these experiments, and their implications, are discussed in section 5. Finally, in section 6, the conclusions which may be drawn from this work are summarized, and possible future directions for research are proposed.

2 Image similarity

As discussed above, any such image database management system requires some measure of image similarity, or, equivalently, the "distance" between

[1] Actually, as pointed out by Picard [16], if there were a unique set of one thousand words which best characterized an image, the query by content problem would be effectively solved.

pairs of images. This distance is frequently the Euclidean distance between points in a multidimensional feature space. No matter what the exact details of the calculation of this distance might be, the important point is that it is a function of some features of the images. In short, the system designer endeavours to select image features and a function of those features so that the distance that results is a measure of *image similarity*.

The aim of an image database system is to assist a human user to retrieve images. In systems which use query by image content, the query is itself an image. Stated simply, the database system computes the distance between the query image and the images in the database and returns to the user those images which are "close" to the query image. This implies that the system's measure of the distance between images corresponds to the user's notion of the dissimilarity of, or perceptual distance between, those images. Moreover, most such systems do not adapt to individual users (with the notable exception of [13]), implying that there is a shared notion of image similarity amongst humans. The implicit assumption of such systems is that there is sufficient overlap between the machine and human measures of image similarity for the database to be useful to all users. In this paper we will attempt to assess the validity of this assumption.

2.1 Human judgment of image similarity

Our approach is to get human subjects to partition a set of colour images with unconstrained content into a number of subsets, with no prompting or guidance. A method for assessing the agreement between pairs of subjects is described in section 3.1, based on statistical measures of reliability [1,3]. The intention of this experiment is to measure the degree to which human subjects' measures of image similarity are consistent. These data will allow us to assess how reasonable it is to expect *any* machine measure of image similarity to agree with a human user. Further analysis of these data might provide insights into desirable properties of clustering algorithms designed to mimic human behaviour.

2.2 Using human similarity assessments to improve machine measures

In any attempt to construct a machine measure of image distance (where a small distance implies a high similarity), there are a great many possible image features from which to choose. Once a set of features is chosen, the designer must then select from a wide variety of techniques for reducing the dimensionality of the feature space, and then computing a distance. Once a distance measure has been decided upon, there is then a large number of clustering techniques from which to choose when organizing the database.

It has been extremely difficult in the past to make an objective assessment of the performance of such systems and, consequently, to make judgments and choices between candidate systems. This is because image retrieval researchers lack large sets of images for which the "ground truth" is known. This is in direct contrast to the situation in the text-based document retrieval community, where much research uses data from the same large, expert-classified datasets, and quantitative comparisons between document retrieval systems are made, notably in the TREC conference series. [2] The documents relevant to a query are decided upon by a panel of experts, so even in the TREC competitions the essentially subjective nature of information retrieval cannot be escaped. Attempts have been made to use the statistics of human subjects' ratings of the performance of vision algorithms to choose between and optimize the algorithms. An example, using human judgments to choose between edge detectors, is found in [7]. Minka and Picard [13] report a system closer in spirit to our work. Their image database system interactively learns groupings of "similar" images from positive and negative examples provided by users during query sessions. However, no explicit attempt is made to measure the agreement between users, or between users and the system.

Our approach is to use a variety of machine systems to cluster the same set of images that was presented to the human subjects into the same number of subsets. It is then possible to compute the degree of agreement between the machine and the human classifiers. Averaged over all humans, this provides a measure of the degree of overlap of each machine measure of image similarity with the common human measure. This measure can thus be used to rank the machine systems, and thus to choose between them.

3 Statistical measures of agreement

3.1 Definition of the agreement measure

Initially, each subject or computer program performs the task described in section 4: the source set of N images is partitioned into M *unlabelled* sub-

[2] **T**ext **RE**trieval **C**onference – The goal of the conference series is to encourage research in information retrieval from large text applications by providing a large test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results. Further information and proceedings are available at: "`http://potomac.ncsl.nist.gov/TREC/`".

sets, some of which may be empty. We wish to compute a measure of the *agreement* between two subjects: we want to measure the *similarity of their image similarity measures.* We will attempt to estimate this by considering their agreement on the similarity or dissimilarity of pairs of images. There is a literature concerning such comparisons, notably in the medical and psychological fields, where the object is typically to measure the degree of agreement between physicians assigning patients into a number of diagnostic categories (for example, see [1,3]). Our problem differs from that paradigm in that the subsets are unlabelled: there is no notion of two subjects assigning images to the "same class".

Since the subsets are undifferentiated, we will consider pairs of images individually. An *agreement measure* can be defined on this basis. Consider the set of images $\Phi = \{I_1, \dots, I_N\}$. Two subjects, A and B, independently partition Φ into M subsets. We will call the resultant sets of subsets of Φ *partitionings* of Φ. These two partitionings are $\Theta_A = \{\theta_{A_1}, \dots, \theta_{A_M}\}$ and $\Theta_B = \{\theta_{B_1}, \dots, \theta_{B_M}\}$. For each pair of images I_i and I_j, there are four possibilities.

Both images are in the same subset in each partitioning:

$$((I_i \in \theta_{A_k}) \wedge (I_j \in \theta_{A_k})) \wedge ((I_i \in \theta_{B_\gamma}) \wedge (I_j \in \theta_{B_\gamma})). \tag{1}$$

Both images are in different subsets in each partitioning:

$$((I_i \in \theta_{A_k}) \wedge (I_j \notin \theta_{A_k})) \wedge ((I_i \in \theta_{B_\gamma}) \wedge (I_j \notin \theta_{B_\gamma})). \tag{2}$$

Images are in the same subset in Θ_A, but different subsets in Θ_B:

$$((I_i \in \theta_{A_k}) \wedge (I_j \in \theta_{A_k})) \wedge ((I_i \in \theta_{B_\gamma}) \wedge (I_j \notin \theta_{B_\gamma})). \tag{3}$$

Images are in different subsets in Θ_A, but the same subset in Θ_B:

$$((I_i \in \theta_{A_k}) \wedge (I_j \notin \theta_{A_k})) \wedge ((I_i \in \theta_{B_\gamma}) \wedge (I_j \in \theta_{B_\gamma})). \tag{4}$$

Equations 1 and 2 describe cases in which subjects A and B have agreed that images I_i and I_j are either similar or dissimilar, and Equations 3 and 4 describe cases of disagreement. We may define a binary variable $X_{ij}(\Theta_A, \Theta_B)$, which is 1 when A and B agree about images i and j, and 0 otherwise. A *raw agreement measure*, $S_{\text{raw}}(\Theta_A, \Theta_B)$, can be obtained by simply counting the number of cases of agreement:

$$S_{\text{raw}}(\Theta_A, \Theta_B) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} X_{ij}(\Theta_A, \Theta_B). \tag{5}$$

The range of $S_{\text{raw}}(\Theta_A, \Theta_B)$ is $[0, \binom{N}{2}]$, so a *normalized agreement measure*, $S(\Theta_A, \Theta_B)$ can be defined as

$$S(\Theta_A, \Theta_B) = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} X_{ij}(\Theta_A, \Theta_B). \tag{6}$$

$S(\Theta_A, \Theta_B)$ provides a measure of agreement between subjects A and B, and is normalized to the range $[0, 1]$, where $S = 0$ indicates complete disagreement, and $S = 1$ complete agreement.

As an example, for the first two human subjects in this study, we obtained $S(\Theta_1, \Theta_2) = 0.8123$ (see section 5.1 for details). At first glance this would seem to indicate a very high level of agreement. This is misleading. We have failed to correct our measure for chance agreements. Any two random partitionings of the image set will have some agreement, which arises purely by chance. If we have a model of the partitioning process, we can compute the expected value of this chance agreement, and use it to correct the agreement measure. Failure to consider chance agreement is a common failing in the image database literature, even though the necessity of such correction is well-known in the field of reliability studies.

Typically, if any comparison with users is made, a percentage intersection of machine and user responses to a query image is given, usually for a small image database. A good example is found in [6], where 131 images were arranged into a two-dimensional grid using a self-organizing map (SOM). Users were asked to select similar images from amongst these in response to query images drawn from the same set. The performance measure quoted was the number X of the user-selected images found within the 24-image square neighbourhood of the query image in the SOM. Let us consider the case in which the images are placed on the grid completely randomly. Neglecting boundary effects (which are also ignored in [6]), if the user selects p relevant images, the expected number $E[X]$ of relevant images within a subset of q images drawn from the remaining $n - 1$ images is

$$E[X] = \frac{pq}{n-1}. \tag{7}$$

In a typical example, the user selected 10 images relevant to the query, giving $E[X] = 1.9$, and eight of these were found within the 24-neighbourhood of the query image. It is misleading to describe this as 80% agreement, as is usually done, since nearly 25% of this performance would be expected if a random arrangement of images were used. It is clear that this effect will be pronounced for the small test databases frequently used for assessing image retrieval systems. As will be seen in the following matter, it is imperative that this be taken into account.

3.2 The kappa statistic

A more appropriate agreement measure is Cohen's kappa statistic [3]:

$$\begin{aligned}\kappa(\Theta_A, \Theta_B) &= \frac{\text{observed agreement} - \text{expected chance agreement}}{1 - \text{expected chance agreement}} \\ &= \frac{S(\Theta_A, \Theta_B) - E[S(\Theta_A, \Theta_B)]}{1 - E[S(\Theta_A, \Theta_B)]}. \end{aligned} \tag{8}$$

The true value of $E[S]$ depends upon subject behaviour. When subsets are labelled, a Bayesian approach is usually adopted. This can be extended to the unlabelled case, as will be discussed in section 3.3. In the absence of such information, we can estimate $E[S]$ by making the minimal assumption that a "blindfolded" subject assigns images to each subset with equal probability. $E[S]$ is then the expected value of S for two independent blindfolded subjects. X_{ij} is now a random variable, and we can find its distribution. We will call this the *blindfolded subject model.*

Consider a pair of images, I_i and I_j. Each image is a member of a subset in each of the partitionings Θ_A and Θ_B. For image I_i, let us label these subsets: $I_i \in \theta_{A_k}$, and $I_i \in \theta_{B_\gamma}$. We then have

$$\begin{aligned}\Pr(I_j \in \theta_{A_k}) &= \frac{1}{M} & \Pr(I_j \in \theta_{B_\gamma}) &= \frac{1}{M} \\ \Pr(I_j \notin \theta_{A_k}) &= \frac{M-1}{M} & \Pr(I_j \notin \theta_{B_\gamma}) &= \frac{M-1}{M}.\end{aligned} \tag{9}$$

Using Equations 1 through 4, we obtain

$$\Pr(X_{ij} = 1) = \frac{1 + (M-1)^2}{M^2} \qquad \Pr(X_{ij} = 0) = \frac{2(M-1)}{M^2}. \tag{10}$$

We can now calculate the statistics of S, since it is the result of $\binom{N}{2}$ trials of a binomial process. We find that

$$E[S] = \frac{1 + (M-1)^2}{M^2} \qquad \sigma^2[S] = \frac{4(M-1)(1 + (M-1)^2)}{N(N-1)M^4}. \tag{11}$$

We can use these in turn to compute the statistics of κ:

$$E[\kappa] = 0 \qquad \sigma^2[\kappa] = \frac{1 + (M-1)^2}{N(N-1)(M-1)}. \tag{12}$$

For the experiment described here, where $N = 100$ and $M = 8$, we find that $E[S] = 0.7813$ and $\sigma[\kappa] = 0.0269$. Applying these new measures for the pair of subjects mentioned in section 3.1, we see that $\kappa(\Theta_1, \Theta_2) = (0.8123 - 0.7813)/(1 - 0.7813) = 0.1420$. Thus the agreement between these two subjects, which at first glance seemed very high, is only 14.20% better than that expected by chance. It is, however, more than 5 standard deviations away from the mean, and thus we can still claim that the agreement between subjects 1 and 2 is significant.

3.3 An alternative agreement measure

It was mentioned in section 3.1 that the blindfolded subject model used above to derive Equation 11 has some short-comings. These arise because of the assumption that the blindfolded subjects assign images to each of subsets with equal probabilities. The nature of this problem becomes obvious if one considers two blindfolded subjects, A and B, who each assign all images to a single subset. For these two subjects $\kappa(AB) = 1$, indicating perfect agreement, and 100% improvement over chance. If, however, these subjects always behave in this way, then $E[S_{\text{raw}}(AB)] = N(N-1)/2$, and $\kappa(AB) = 1$ *always.* Even though these subjects exhibit "perfect agreement", this arises entirely from their prior bias, and conveys no information about their judgments of image similarity.

Preliminary experiments indicated that human subjects and, even more markedly, computer classification algorithms did indeed produce subsets of greatly varying sizes. For example, the computer algorithm referred to as X_2 in section 5 produced subsets of sizes {6, 1, 8, 6, 47, 23, 4, 5}. Human subjects did not exhibit such large bias towards particular subsets, although they too did not produce subsets of equal sizes (see section 5.2 for details). The existence of this bias means that the blindfolded subject model will not account well for the observed data, and an alternative is required.

This problem can be avoided by taking into account the frequency with which subjects assign images to each subset. This is easiest to understand if the partitioning data are presented as a matrix $\mathbf{A}$ of intersections between subsets. Table 1 shows an hypothetical example with $N = 20$ and $M = 3$.

The row and column marginal sums from Table 1 can be used to estimate each subject's probability of assigning an image to a given subset. This "Bayesian" expected value for element a_{ij} is [3]:

$$E_B[a_{ij}] = \frac{a_{i+}a_{+j}}{N}. \qquad (13)$$

The matrix of expected intersections is shown in Table 2.

Table 1 contains enough information to calculate $S_{\text{raw}}(AB)$, as defined in

		Subject A			
		θ_{A_1}	θ_{A_2}	θ_{A_3}	a_{+j}
Subject B	θ_{B_1}	8	2	1	11
	θ_{B_2}	2	3	0	5
	θ_{B_3}	2	1	1	4
	a_{i+}	12	6	2	20

Table 1
Intersections between subsets created by subjects A and B.

		Subject A		
		θ_{A_1}	θ_{A_2}	θ_{A_3}
Subject B	θ_{B_1}	6.6	3.3	1.1
	θ_{B_2}	3.0	1.5	0.5
	θ_{B_3}	2.4	1.2	0.4

Table 2
Expected intersections between subsets given frequency data for subjects A and B.

Equation 5. If a_{ij} is the element in the ith row and jth column, then

$$S_{\text{raw}}(AB) = \sum_{1 \le i,j \le M} \binom{a_{ij}}{2} + \frac{1}{2} \sum_{\substack{1 \le k,l,m,n \le M \\ k \ne m,\, l \ne n}} a_{kl} a_{mn}, \tag{14}$$

where the usual definitions $\binom{1}{2} \equiv 0$ and $\binom{0}{2} \equiv 0$, are made. The first sum in Equation 14 counts agreements of the form specified by Equation 1 (agreement that two images are similar), and the second sum counts the agreements corresponding to Equation 2 (agreement that the images are dissimilar). The factor of 1/2 in the second term arises because each pair of matrix elements appears twice in the sum.

3.3.1 Estimating expected agreement given prior subset probabilities

This formulation provides us with a means of obtaining an approximation to the expected value of S_{raw}. Although an exact calculation of $E[S_{\text{raw}}]$ is possible, it requires the enumeration of all possible assignments of images to subsets, since S_{raw} is a nonlinear function of the $\{a_{ij}\}$. Rather than performing this time-consuming calculation for each of the 630 pairs of partitionings considered, we will approximate $E[S_{\text{raw}}]$ by

$$\tilde{E}[S_{\text{raw}}] = S_{\text{raw}}(\{E[a_{ij}]\}), \tag{15}$$

the value of S_{raw} computed from the matrix of expected intersections. It is necessary to scale the values to integers so that $\binom{a_{ij}}{2}$ can be computed.[3] Since the $\{a_{ij}\}$ are always rational, and represent frequency information, this can always be done. The value obtained can then be normalized to obtain $\tilde{E}_B[S]$ (analogous to $E[S]$ in Equation 11), which is used to calculate a new statistic, κ_B, based on this new estimate of the expected chance agreement.

For the data in Table 2 we obtain $\tilde{E}_B[S] = 0.5063$. This compares with $E[S] = 0.5556$ from the blindfolded subject model. The corresponding agreement measures are $\kappa = -0.0066$ and $\kappa_B = 0.0937$. Taking into account the subjects' biases towards the various subsets yields a κ_B indicating better than chance agreement, whereas the blindfolded subject model gives a result worse than chance. Indeed, the κ_B statistic seems to be always more "forgiving" than the blindfolded subject model. The disadvantage of κ_B is that the expected agreement is now dependent on the pair of subjects being considered, making the interpretation of large tables of data more difficult.

Simulations were performed to check the validity and possible bias of this estimate of $E_B[S_{\text{raw}}]$. Pairs of partitionings from the experiments described in section 4 were used to provide probabilities for subsets, and pairs of partitionings were created in which elements were assigned to the subsets according to these probabilities. The similarity between these partitionings was computed, and the results accumulated over many repetitions of the experiment. The average similarity obtained from this simulation could then be compared with the estimate provided by Equation 15 for the original pair of subjects.

Table 3 shows the results for human subjects S_{13} and S_{16} and for machine algorithms X_3 and X_{16}, after 10 repetitions of comparisons between 10000 partitionings made with their subset probabilities. It can be seen that in neither case is the average difference between the estimate and the value obtained by simulation greater than its standard deviation. From this, and numerous similar tests, we can conclude that Equation 15 is an accurate, unbiased estimator of $E_B[S_{\text{raw}}]$.

3.3.2 Dependence of $E_B[S]$ on subset probability distributions

As mentioned above, κ_B is a more "forgiving" statistic than the κ, since the $E_B[S]$ calculated from the experimentally observed subset frequencies is less than that predicted by the blindfolded subject model. The chance-corrected agreement is thus greater. It is interesting to investigate the value of $E_B[S]$ for a variety of subset probability distributions. In order to do this, a para-

[3] Alternatively, the extension of the binomial coefficient function to non-integer arguments in terms of gamma functions could be used. A description of the properties of this function may be found in [5].

S_{13} and S_{16}			X_3 and X_{16}		
Simulation	Calculated	Difference	Simulation	Calculated	Difference
0.747813	0.747852	-0.000039	0.622178	0.622240	-0.000062
0.747755	0.747852	-0.000097	0.622506	0.622240	0.000266
0.747798	0.747852	-0.000053	0.622655	0.622240	0.000416
0.747870	0.747852	0.000019	0.622409	0.622240	0.000169
0.747819	0.747852	-0.000033	0.622265	0.622240	0.000025
0.747938	0.747852	0.000087	0.621858	0.622240	-0.000382
0.747843	0.747852	-0.000008	0.622473	0.622240	0.000233
0.747930	0.747852	0.000079	0.622124	0.622240	-0.000116
0.747999	0.747852	0.000147	0.622162	0.622240	-0.000078
0.747930	0.747852	0.000078	0.622391	0.622240	0.000151
avg. diff. $\pm\sigma$: 0.000018 ± 0.000073			average diff. $\pm\sigma$: 0.000062 ± 0.000219		

Table 3
Comparison between $E_B[S_{\mathrm{raw}}]$ estimated by simulation, and that given by Equation 15 for subjects S_{13} and S_{14}, and machine algorithms X_3 and X_{16}.

metric probability density function for subset membership is required, which allows the subset probabilities to vary, under the control of a parameter k, from uniform to a delta function for a single subset. Such a function can be constructed from an exponential distribution. On the interval $[0,1]$,

$$\Pr(x) = \frac{ke^{-kx}}{1-e^{-k}} \tag{16}$$

has the required properties. As $k \to 0$, $\Pr(x)$ approaches the uniform distribution. As $k \to \infty$, $\Pr(x)$ approaches a delta function at the origin. If the interval $[0,1]$ is divided into M equal subintervals, labelled $\{0, 1, \ldots, i, \ldots, M-1\}$, Equation 16 can be integrated to provide the probability of a uniformly-distributed x falling into each subinterval:

$$\Pr(x \in i) = \frac{e^{-\frac{ki}{M}}(1-e^{-\frac{k}{M}})}{1-e^{-k}}. \tag{17}$$

Pairs of 8-subset partitionings were produced, each having the same subset probability distributions according to Equation 17. The agreement between these partitionings was computed. This simulation was repeated 1000 times for each value of k, so that $E_B[S]$ as a function of k could be estimated. The resultant curve appears in Figure 1, where k appears on a log scale.

As expected, as $k \to 0$, $E_B[S] \to \frac{1+(M-1)^2}{M^2}$, the value derived from the blind

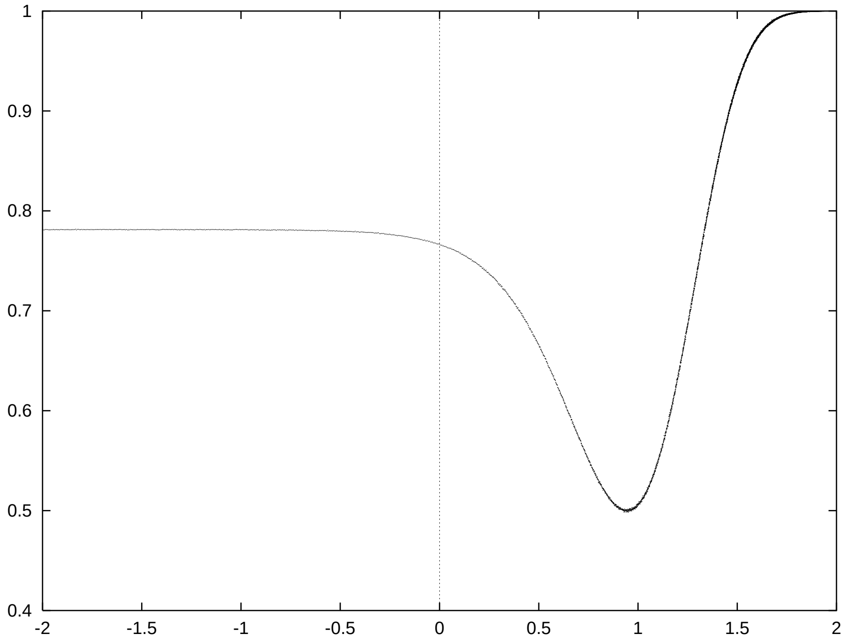

Fig. 1. Estimated $E_B[S]$ as a function of $\log_{10}(k)$.

subject model. As $k \to \infty$, $E_B[S] \to 1$, which corresponds to the degenerate case described in section 3.3, in which all images are assigned to a single subset. The values of $E_B[S]$ (as estimated by Equation 15) for the distributions observed in the experiments reported in section 5 ranged from 0.576633 to 0.767303, indicating that in all cases the distributions fell into the region of Figure 1 in which $E_B[S]$ is less than the blind subject model value. This explains why the κ_B is more forgiving than κ in the experiments reported in section 5.

4 Experiments

The task that the human subjects were asked to perform was to partition a set of N images into at most M subsets. This task was performed using a computer program running on SUN workstations, which presented each subject initially with a source image set, and M empty sets. Images could be dragged from any image set and dropped in another image set using the mouse. When all images from the source image set had been assigned to subsets, the partitioning could be saved.

For this particular experiment, the source image set consisted of 100 colour images selected at random from a set of 500 unconstrained images provided by Télévision Suisse Romande. [4] The set of 500 images contained some "runs"

[4] The 500 images constituted a consecutive subset of the full 10,000 images provided

of images from the same video footage, and thus some highly similar images could be expected. The images were randomly ordered so that images from the same run would not necessarily be presented to the subject adjacently. The same set of 100 images was used for all experiments reported here. Sample images are shown in Figure 2.

Fig. 2. Sample images from the source set of 100 images (originals were in colour).

Subjects were given a brief demonstration of how the program worked, and told that the notion of image similarity was entirely their choice. The task was performed by 10 members of the computer vision research group at the Université de Genève (who may be considered to be have some expert knowledge), and by 8 undergraduate students and lay-people.

The same set of images was classified into a binary tree using Ascendant Hierarchical Classification (AHC) [2], using a variety of distance measures derived by applying Correspondence Analysis (CA) [10], Principal Components Analysis (PCA) and Normalized Principal Components Analysis (NPCA) to a range of colour, segment and arc statistics, as described in detail in [17,12]. For each factor analysis technique, the classification was performed using all available features, and also with the features relating to colour information excluded. For each of these cases, a classification was done using 2, 4 and all of the ranked factors. There were thus 18 different machine partitionings of the images. The third level of a binary tree contains at most 8 classes, and is thus comparable with the human classifications described above.

5 Results and discussion

5.1 *Agreement between humans*

Table A.1 shows the agreement between all pairs of human subjects, as indicated by the κ values. For each subject S_i, $\mu[\kappa_{S_i}]$ is the average agreement

by Télévision Suisse Romande.

with all other subjects, excluding themselves:

$$\mu[\kappa_{S_i}] = \frac{1}{N_{\mathrm{subjects}}} \sum_{\substack{1 \le j \le N_{\mathrm{subjects}} \\ j \ne i}} \kappa(\Theta_i, \Theta_j). \tag{18}$$

Extreme values, some of which are discussed below, are highlighted with a shaded background.

Summary statistics of these data are shown in Table 4. These statistics were calculated with the self-agreement of 1 for each subject excluded. The most important result is that the average agreement between human subjects is 18.53% better than would be expected by chance from the blindfolded subject model. Whilst perhaps lower than some might have expected, this is still more than 6 standard deviations away from the model mean. In fact, given the model, $\Pr(\kappa \ge 0.1853) = 9.202 \times 10^{-13}$. The hypothesis that the subjects are "blindfolded" may be rejected with virtually complete confidence.

In order to make a rough assessment of the effect of a degree of expertise with image processing and computer vision, the statistics were also calculated separately for the 10 members of the computer vision group ("experts"), and the 8 lay-people.

	mean	median	std. dev.	κ min.	κ max.
All Subjects	0.1853	0.1919	0.1241	-0.1627	0.4708
Experts	0.2368	0.2335	0.0972	0.0294	0.4708
Lay People	0.1359	0.1300	0.1604	-0.1627	0.4246

Table 4
Statistics summarizing the agreement between human subjects, using κ. Full data are found in Table A.1.

These data show that the agreement between experts is significantly higher than that between lay people, and that the variation between experts is less than that between lay people. This result would seem to indicate that the human image similarity measure is partially learnt.

Another interesting observation is that the subject for whom the average agreement with all other subjects is least is H_3. Subject H_3 is a colour-blind male. This result hints, not unexpectedly, that colour information is important in human judgments of image similarity, and that purely morphological features would prove an inadequate basis for a machine image similarity measure. A single result, however, does not provide a sound statistical basis for this conclusion.

The data used to construct Table A.1 were also analyzed using the κ_B statistic. The results appear in Table A.2, and the summary in Table 5.

	mean	median	std. dev.	κ_B min.	κ_B max.
All Subjects	0.3450	0.3381	0.0926	0.1736	0.6266
Experts	0.3773	0.3625	0.0822	0.2225	0.5868
Lay People	0.3181	0.2781	0.1128	0.1736	0.6266

Table 5
Statistics summarizing the agreement between human subjects, using κ_B. Full data are found in Table A.2.

Using κ_B causes some reordering of the average agreement measure for the subjects. The average change in rank for the subjects was 2.1. Despite this, the trends remain the same. The agreement between experts is greater than that between lay people, and the variance is smaller. The overall variance is less than for the κ statistic, which is to be expected, since the κ_B statistic is calculated using an expected agreement value tailored specifically to the particular pair of subjects being considered.

5.2 Agreement between machine partitionings

Analysis of the agreement between machine techniques can give an indication of the variance between a number of methods, and also of the stability of each technique under changes in its parameters. As discussed in section 4, the images were partitioned using AHC. The distance measures upon which these clusterings were based were given by a variety of different factor analysis methods, either with or without colour information, and with differing numbers of retained factors. Table 6 provides a key to the labelling of these machine techniques in the tables that follow.

X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9
CA c 2	CA c 4	CA c all	CA nc 2	CA nc 4	CA nc all	NPCA c 2	NPCA c 4	NPCA c all
X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}
NPCA nc 2	NPCA nc 4	NPCA nc all	PCA c 2	PCA c 4	PCA c all	PCA nc 2	PCA nc 4	PCA nc all

Table 6
Key to computer partitioning methods. “c” means that colour features were used, “nc” means that they were not. The last field indicates how many factors were retained for the clustering. Details of these algorithms and features may be found in [17].

Table 7 provides a summary of the agreement between the various machine partitionings, as measured by κ. The complete data appear in Table A.3. Both tables are divided into sections so that the variance within and between the various factor analysis techniques can be seen more easily, and extreme values are again indicated by shading.

	mean	median	std. dev.	κ min.	κ max.
All Variations	-0.2662	-0.3142	0.3219	-0.7805	0.9511
All CA	-0.2190	-0.6651	0.6394	-0.7805	0.9511
CA colour	0.3018	-0.0094	0.4592	-0.0362	0.9511
CA no colour	0.6996	0.6565	0.1063	0.5964	0.8458
All NPCA	0.2863	-0.0777	0.1856	-0.3059	0.3702
NPCA colour	-0.0971	-0.1858	0.2159	-0.3059	0.2002
NPCA no colour	0.1996	0.1790	0.1317	0.0497	0.3702
All PCA	0.0413	-0.2375	0.4181	-0.3603	0.7433
PCA colour	0.7100	0.6999	0.0241	0.6869	0.7433
PCA no colour	0.3480	0.3258	0.0993	0.2390	0.4791

Table 7
Statistics summarizing the agreement between machine partitionings, using κ. Full data are found in Table A.3.

The first thing to note is that the average agreement between machine techniques over all the used permutations of factor analysis method, feature choice and number of factors retained is -0.2662. The agreement between machine partitionings is *worse* than that which would be expected if the machines simply assigned images to subsets with uniform probability.

At first, this seems to be a very depressing result for proponents of these methods of unsupervised image clustering. It must be remembered, however, that the behaviour of these techniques does not correspond at all well to the blindfolded subject model. All the machine clusterings exhibit a strong tendency to produce one or two very large subsets, and several very small ones. This can be explained by the fact that the distribution of values of each factor is approximately bell-shaped. There is thus a large number of images with values close to the means of the factors, and only a few images in the tails of the distribution. This results, using AHC, in several large subsets close to the mean values of the factors, and other small subsets containing the remaining images.

Inspection of the machine partitionings reveals that the mean maximum subset size was 34, with standard deviation 8.4. The mean minimum subset size was 1.7 with standard deviation 1.2. This compares with a mean maximum of 24 with standard deviation of 4.7, and a mean minimum of 4.5 with standard deviation of 2.0 for the human subjects. These results indicate that the machine techniques deviate from the uniform subset probability assumed by the blindfolded user model even more than the human subjects. This suggests that a clustering technique which produced subsets of more uniform size than AHC would better match human performance. Perhaps using a 1-metric, rather than the standard Euclidean 2-metric, would help this.

This bias means that the κ_B statistic is a better indicator of improvement over chance performance for machine techniques. Table 8 shows the summary of these data reanalyzed using κ_B. The full data are in Table A.4.

	mean	median	std. dev.	κ_B min.	κ_B max.
All Variations	0.2023	0.1574	0.1757	0.0176	0.9743
All CA	0.3404	0.0848	0.3437	0.0675	0.9743
CA colour	0.6185	0.4508	0.2518	0.4303	0.9743
CA no colour	0.8364	0.8134	0.0605	0.7765	0.9192
All NPCA	0.2862	0.2537	0.0949	0.2026	0.5511
NPCA colour	0.3019	0.2569	0.1037	0.2036	0.4453
NPCA no colour	0.4116	0.3518	0.0990	0.3319	0.5511
All PCA	0.3144	0.0877	0.2987	0.0626	0.8014
PCA colour	0.7824	0.7780	0.0141	0.7677	0.8014
PCA no colour	0.5517	0.5202	0.0673	0.4897	0.6453

Table 8
Statistics summarizing the agreement between machine partitionings, using κ_B. Full data is found in Table A.4.

Now that the non-uniform probability of assignment to subsets has been taken into account, it emerges that the agreement between all these techniques is better than would be expected by chance. CA has the greatest self-agreement under changes of the number of factors retained and NPCA the least, but the values are within one standard deviation of each other. Few conclusions can be drawn about the relative stability of these techniques. A self-agreement of 0.3404 seems very low for a completely deterministic technique.

From Table A.4 we can see that the self-agreement of these techniques is very much higher when considering only the retention of 4 factors against the same technique with all factors. For CA, for example, it is 0.9743 when colour features are used, and 0.9192 when they are not. This confirms that the data are very well represented by only the first four factors, as should be the case (the first four factors explain 88.35% of the variance in the CA space when colour features are included).

The two techniques with the lowest agreement are X_3 and X_{10}: CA with colour features and all factors retained against NPCA, with no colour information and only two factors.

5.3 Agreement between human and machine partitionings

It was established in section 5.2 that both human and machine partitionings exhibit such bias towards large subsets that the blindfolded user model greatly overestimates the expected chance agreement. This means that κ_B is the more appropriate measure of agreement. It is nevertheless interesting to highlight some observations drawn from the κ statistic. In summary, the mean κ agreement between all machine partitionings and all human subjects was -0.2806. The best technique, averaged over the six variants, was PCA, with -0.1844. The worst was CA, with -0.4663. Averaged over the three different numbers of retained factors, the best technique was PCA with colour features, with $\mu_{\mathrm{PCA_c}}[\kappa] = -0.1393$, and the worst was CA with no colour, with $\mu_{\mathrm{CA_{nc}}}[\kappa] = -0.5060$. The overall standard deviation was 0.1648, so these results are significant.

It is interesting to note that the *only* human subject with positive κ values with any machine technique was H_{16}, for whom the κ values with $\{X_{13}, X_{14}, X_{15}\}$ (PCA with colour features) were $\{0.0876, 0.0266, 0.0109\}$ respectively. Subject H_{16} is in fact the first-named author of this paper, who has worked with these machine image clustering methods for 12 months. This result has two implications. First, it points to the dangers of allowing the creators of image similarity measures to be their sole assessors. Secondly, and positively, it would seem to indicate that human measures of image similarity are partially learnt: after prolonged interaction with an image database system, the human begins to judge image similarity in a similar way. The ability of humans to adapt to a system which they are using relaxes the demands on the system.

Table 9 shows the summary of the κ_B agreement between all human and machine partitionings of the images. The full data are found in Table A.5, where extreme values are highlighted by shading.

The first thing to note is that, using κ_B, in all cases the agreement between machine and human partitionings is positive. Secondly, the difference between individual machine methods, when average across all human users, is much less than it was when using κ, indicating that the differing subset sizes were largely responsible for those disparities. The machine technique with the least improvement over chance was X_6 (CA, no colour features, all factors retained), with $\mu[\kappa_B] = 0.0573$. The best was X_9 (NPCA, colour features, all factors retained), with $\mu[\kappa_B] = 0.1477$. The standard deviation over all machine clustering variants was 0.0240, so this difference is significant.

When averaged over all 6 variants for each technique, NPCA was best, with

	mean	median	std. dev.	κ_B min.	κ_B max.
All Variations	0.1067	0.1058	0.0338	0.0250	0.2312
All CA	0.0915	0.0836	0.0371	0.0250	0.1825
CA colour	0.1169	0.1186	0.0308	0.0683	0.1825
CA no colour	0.0662	0.0621	0.0226	0.0250	0.1229
All NPCA	0.1233	0.1235	0.0281	0.0589	0.1911
NPCA colour	0.1388	0.1400	0.0235	0.0817	0.1911
NPCA no colour	0.1078	0.1068	0.0233	0.0589	0.1633
All PCA	0.1052	0.1007	0.0276	0.0553	0.2312
PCA colour	0.1167	0.1086	0.0296	0.0719	0.2312
PCA no colour	0.0936	0.0914	0.0197	0.0553	0.1328

Table 9
Statistics summarizing the agreement between human and machine partitionings, using κ_B. Full data are found in Table A.5.

0.1233, and CA was worst, with 0.0915, with a standard deviation of 0.0130. Again, this would seem to be significant.

For each factor analysis technique, the use of colour features gave improved agreement with the human subjects, corroborating the conjecture in section 5.1.

For CA and PCA, performance with colour features was best when only two factors were retained, but for NPCA, performance was best when all factors were retained. In none of these cases, however, was the difference greater than one standard deviation, so the computational savings made by using a reduced number of factors could perhaps be justified.

Over all these techniques, however, the greatest κ_B with any subject was 0.2312, between X_{13} and S_{16} (again, the first-named author). This compares with 0.6266 between H_4 and H_{18}. The average κ_B between humans and humans was 0.3464, whereas between humans and machine partitionings it was 0.1067. The machine techniques reviewed here do provide significantly better than chance agreement with human subjects, but are a long way from being as good as the “average” human.

6 Conclusion

In this paper we have shown that a rigorous assessment of the agreement between two partitionings of a set of images into unlabelled subsets is possible. The most important feature of any such measure of agreement is that it *must* take into account the agreement expected by chance. We believe that reports of the performance of image database systems have consistently failed to do this. The actual value of the expected chance agreement depends on the model of user behaviour selected. We have shown that the simplest possible model, which assumes equal probabilities for each subset, does not describe the measured behaviour of human subjects or machine partitioning techniques adequately. We have thus proposed the κ_B statistic, which explicitly takes into account the subset probabilities for each subject.

Using this statistic, we have found that the agreement between human subjects is on average 34.64% higher than would be expected by chance, almost 4 standard deviations away from the expected value. This indicates that there is indeed some shared notion of image similarity, but the value is still a long way from 100% agreement. There is also much variance in the agreement between pairs of human subjects: different subjects agree in different ways.

These observations suggest that a truly successful image database system should attempt to model the individual user, so that the image distance measure is at least partially learnt. In practice, it is likely that the appropriate measure will depend not only on the individual user, but also on the genre of images, and the task which the user is performing.

There is evidence that the human notion of image similarity is at least partially learnt, as demonstrated by the higher agreement between experts than between lay people. Also, it seems that humans can adapt to an image classification system, and subconsciously adjust their notions of image similarity accordingly. This suggests that the user will be able to adapt to a image database system, as well as *vice versa*.

Measurements of the agreement between human subjects and a variety of differing machine image partitioning techniques showed that the agreement between machine methods and humans was much less than that between pairs of humans. Despite this, the collection of human image clusterings provided an objective means of assessing the performance of the image clustering systems. It was possible to conclude that colour features should definitely be used, and that Normalized Principal Components Analysis was the best of the factor analysis techniques tried. More importantly, the utility of the methodology has been demonstrated. Larger scale experiments, especially in terms of the number of human subjects assessed, would be necessary to obtain clearly significant distinctions between competing measures of image similarity.

Future work could include experiments in which subjects were asked to assign images to classes which had had initial seed images assigned to them. This would allow an assessment using the previously established methodology for labelled classes. Such an approach would also be more appropriate for assessing the efficacy of a similarity measure for responding to query images.

Finally, we reiterate our belief in the importance of gathering such ground truth for assessing the performance of image database systems. In this paper we have demonstrated a methodology for analyzing and applying such data, and, in particular, shown the importance of using statistics which consider chance agreements.

Acknowledgement

The authors would like to thank Marianne Gex-Fabry for her advice on measures of agreement and reliability.

References

[1] John J. Bartko and William T. Carpenter. On the methods and theory of reliability. *The Journal of Nervous and Mental Disease*, 163(5):307–317, 1976.

[2] E. Diday, G. Govaert, Y. Lechevallier, and J. Sidi. Clustering in pattern recognition. In J.-C. Simon and R. Haralick, editors, *Digital Image Processing*, pages 19–58. D. Reidel Publishers, 1981.

[3] Graham Dunn. *Design and analysis of reliability studies; the statistical evaluation of measurement errors.* Oxford University Press, 200 Madison Avenue, New York, NY 10016, 1989.

[4] Myron Flickner, Harpreet Sawhney, Wayne Niblack, Jonathon Ashley, Qian Huang, Byron Dom, Monika Gorkani, Jim Hafner, Denis Lee, Dragutin Petkovic, David Steele, and Peter Yanker. Query by image and video content: The QBIC system. *IEEE Computer*, 28(9):23–32, September 1995.

[5] David Fowler. The binomial coefficient function. *American Mathematical Monthly*, 103(1):1–17, January 1996.

[6] K.A. Han and S.-H. Myaeng. Image organization and retrieval with automatically constructed feature vectors. In H.-P. Frei, D. Harman, P. Schäuble, and R. Wilkinson, editors, *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in*

Information Retrieval (SIGIR'96), pages 157–165, Zürich, Switzerland, August 1996.

[7] Mike Heath, Sudeep Sarkar, Thomas Sanocki, and Kevin Bowyer. Comparison of edge detectors: A methodology and initial study. In *Proceedings of the 1996 Conference on Computer Vision and Pattern Recognition (CVPR '96)*, pages 143–148. IEEE Computer Society Press, June 1996.

[8] Anil K. Jain and Aditya Vailaya. Image retrieval using color and shape. *Pattern Recognition*, 29(8):1233–1244, August 1996.

[9] Toshikazu Kato. Database architecture for content-based image retrieval. *SPIE Image Storage and Retrieval Systems*, 1662:112–123, February 1992.

[10] Ludovic Lebart, Alain Morineau, and Jean-Pierre Fénelon. *Traitement des données statistiques; méthodes et programmes.* Dunod, Paris, 1979.

[11] Michael S. Lew, D. P. Huijsmans, and Dee Denteneer. Content based image retrieval: KLT, projections, or templates. In A. W. M. Smeulders and R. Jain, editors, *Image Databases and Multi-Media Search*, pages 27–34. Amsterdam University Press, August 1996.

[12] Ruggero Milanese, David McG. Squire, and Thierry Pun. Correspondence analysis and hierarchical indexing for content-based image retrieval. In P. Delogne, editor, *IEEE International Conference on Image Processing*, volume 3, pages 859–862, Lausanne, Switzerland, September 1996.

[13] T. P. Minka and R. W. Picard. Interactive learning using a "society of models". In *Proceedings of the 1996 Conference on Computer Vision and Pattern Recognition (CVPR '96)*, pages 447–452. IEEE Computer Society Press, June 1996.

[14] Farzin Mokhtarian and Sadegh Abbasi Josef Kittler. Efficient and robust retrieval by shape content through curvature scale space. In A. W. M. Smeulders and R. Jain, editors, *Image Databases and Multi-Media Search*, pages 35–42. Amsterdam University Press, August 1996.

[15] A. Pentland, R. W. Picard, and S. Sclaroff. Photobook: Tools for content-based manipulation of image databases. *International Journal of Computer Vision*, 13(3), June 1996.

[16] Rosalind W. Picard. Toward a visual thesaurus. Technical Report 358, MIT Media Laboratory Perceptual Computing Section, 20 Ames St., Cambridge MA 02139, 1995.

[17] Thierry Pun and David McG. Squire. Statistical structuring of pictorial databases for content-based image retrieval systems. *Pattern Recognition Letters*, 17:1299–1310, 1996.

[18] Stan Sclaroff. Encoding deformable shape categories for efficient content-based search. In A. W. M. Smeulders and R. Jain, editors, *Image Databases and Multi-Media Search*, pages 107–114. Amsterdam University Press, August 1996.

[19] Aya Soffer and Hanan Samet. Handling multiple instances of symbols in pictorial queries by image similarity. In A. W. M. Smeulders and R. Jain, editors, *Image Databases and Multi-Media Search*, pages 51–58. Amsterdam University Press, August 1996.

A Tables of Experimental Results

	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_9	H_{10}	H_{11}	H_{12}	H_{13}	H_{14}	H_{15}	H_{16}	H_{17}	H_{18}
H_1	1.0000	0.1420	0.1393	0.0737	0.2390	0.2335	0.0986	0.3785	0.1180	0.1891	0.2196	0.2547	0.3332	0.2603	0.1448	0.2880	0.2159	0.0876
H_2	0.1420	1.0000	0.1328	0.0155	0.1273	0.1513	0.1014	0.1654	0.0543	0.1568	0.2095	0.1337	0.0756	0.1079	0.0571	0.2002	0.2132	-0.1627
H_3	0.1393	0.1328	1.0000	-0.0888	0.0599	0.0303	0.0174	0.0165	0.0091	0.1005	0.0903	0.0682	0.1116	0.0608	-0.0011	0.1217	0.0737	-0.1544
H_4	0.0737	0.0155	-0.0888	1.0000	0.3323	0.0885	0.1642	0.1688	0.0008	0.0959	0.2833	0.1928	0.2750	0.2483	0.4653	-0.0122	0.2390	0.4246
H_5	0.2390	0.1273	0.0599	0.3323	1.0000	0.2132	0.3942	0.2713	0.1956	0.4182	0.2769	0.3415	0.3868	0.3674	0.3240	0.2178	0.4283	0.1707
H_6	0.2335	0.1513	0.0303	0.0885	0.2132	1.0000	0.1836	0.2473	0.0294	0.2298	0.1919	0.2880	0.1781	0.1587	0.1873	0.2141	0.2547	-0.0122
H_7	0.0986	0.1014	0.0174	0.1642	0.3942	0.1836	1.0000	0.0959	0.0793	0.2630	0.2233	0.2843	0.2279	0.1956	0.1725	0.1457	0.2843	-0.0436
H_8	0.3785	0.1651	0.0165	0.1688	0.2713	0.2473	0.0959	1.0000	0.2076	0.2030	0.2372	0.1891	0.3489	0.4071	0.2381	0.1965	0.2483	0.0848
H_9	0.1180	0.0543	0.0091	0.0008	0.1956	0.0294	0.0793	0.2076	1.0000	0.1384	0.1411	0.0211	0.1679	0.2150	0.0515	0.0451	0.1965	-0.0630
H_{10}	0.1891	0.1568	0.1005	0.0959	0.4182	0.2297	0.2630	0.2030	0.1384	1.0000	0.1956	0.2769	0.2372	0.3766	0.1282	0.2455	0.4468	-0.0584
H_{11}	0.2196	0.2095	0.0903	0.2833	0.2769	0.1919	0.2233	0.2372	0.1411	0.1956	1.0000	0.1707	0.3397	0.2797	0.4708	0.1670	0.3868	0.1014
H_{12}	0.2547	0.1337	0.0682	0.1928	0.3415	0.2880	0.2843	0.1891	0.0211	0.2769	0.1707	1.0000	0.1513	0.2519	0.1624	0.2316	0.3406	0.0423
H_{13}	0.3332	0.0756	0.1116	0.2750	0.3869	0.1781	0.2279	0.3489	0.1679	0.2372	0.3397	0.1513	1.0000	0.2676	0.3683	0.2603	0.3175	0.2353
H_{14}	0.2603	0.1079	0.0608	0.2483	0.3674	0.1587	0.1956	0.4071	0.2150	0.3766	0.2797	0.2519	0.2676	1.0000	0.3046	0.1448	0.3794	0.1254
H_{15}	0.1448	0.0571	-0.0011	0.4653	0.3240	0.1873	0.1725	0.2381	0.0515	0.1282	0.4708	0.1624	0.3683	0.3046	1.0000	0.0848	0.2990	0.2926
H_{16}	0.2880	0.2002	0.1217	-0.0122	0.2178	0.2141	0.1457	0.1965	0.0451	0.2455	0.1670	0.2316	0.2603	0.1448	0.0848	1.0000	0.2335	-0.0205
H_{17}	0.2159	0.2132	0.0737	0.2390	0.4283	0.2547	0.2843	0.2483	0.1965	0.4468	0.3868	0.3406	0.3175	0.3794	0.2990	0.2335	1.0000	0.1162
H_{18}	0.0876	-0.1627	-0.1544	0.4246	0.1707	-0.0122	-0.0436	0.0848	-0.0630	-0.0584	0.1014	0.0423	0.2353	0.1254	0.2926	-0.0205	0.1162	1.0000
$\mu(\kappa_{H_i})$	0.2009	0.1106	0.0463	0.1745	0.2803	0.1686	0.1699	0.2179	0.0946	0.2143	0.2344	0.2001	0.2519	0.2442	0.2206	0.1626	0.2749	0.0686

Table A.1
Agreement between human subjects on the image partitioning task, as measured by the κ statistic.

	H_1	H_2	H_3	H_4	H_5	H_6	H_7	H_8	H_9	H_{10}	H_{11}	H_{12}	H_{13}	H_{14}	H_{15}	H_{16}	H_{17}	H_{18}
H_1	1.0000	0.2693	0.2799	0.3076	0.3408	0.3475	0.2252	0.5041	0.3083	0.2817	0.3467	0.3775	0.4457	0.3925	0.3201	0.3631	0.3058	0.3523
H_2	0.2693	1.0000	0.2771	0.2661	0.2469	0.2803	0.2305	0.3361	0.2607	0.2561	0.3407	0.2790	0.2342	0.2699	0.2528	0.2876	0.3062	0.1766
H_3	0.2799	0.2771	1.0000	0.1991	0.2034	0.1921	0.1736	0.2298	0.2367	0.2214	0.2541	0.2375	0.2763	0.2438	0.2187	0.2327	0.1986	0.1920
H_4	0.3076	0.2661	0.1991	1.0000	0.4960	0.3208	0.3727	0.4100	0.2999	0.3061	0.4726	0.4070	0.4693	0.4546	0.6215	0.2176	0.4162	0.6266
H_5	0.3408	0.2469	0.2034	0.4960	1.0000	0.3214	0.4722	0.4120	0.3622	0.4772	0.3870	0.4430	0.4838	0.4741	0.4565	0.2901	0.4866	0.4062
H_6	0.3475	0.2803	0.1921	0.3208	0.3214	1.0000	0.3013	0.4017	0.2415	0.3208	0.3264	0.4077	0.3195	0.3118	0.3563	0.3003	0.3432	0.2835
H_7	0.2252	0.2305	0.1736	0.3727	0.4722	0.3013	1.0000	0.2752	0.2746	0.3433	0.3464	0.3990	0.3548	0.3360	0.3391	0.2312	0.3625	0.2565
H_8	0.5041	0.3361	0.2298	0.4100	0.4120	0.4017	0.2752	1.0000	0.4134	0.3443	0.4024	0.3659	0.4930	0.5431	0.4293	0.3335	0.3819	0.3793
H_9	0.3083	0.2607	0.2367	0.2999	0.3622	0.2415	0.2746	0.4134	1.0000	0.3040	0.3381	0.2470	0.3625	0.4046	0.3000	0.2225	0.3513	0.2873
H_{10}	0.2817	0.2561	0.2214	0.3061	0.4772	0.3208	0.3433	0.3443	0.3040	1.0000	0.3033	0.3751	0.3442	0.4710	0.2855	0.2982	0.4911	0.2311
H_{11}	0.3467	0.3407	0.2541	0.4726	0.3870	0.3264	0.3464	0.4024	0.3381	0.3033	1.0000	0.3210	0.4619	0.4197	0.5868	0.2718	0.4692	0.3708
H_{12}	0.3775	0.2790	0.2375	0.4070	0.4430	0.4077	0.3990	0.3659	0.2470	0.3751	0.3210	1.0000	0.3097	0.3986	0.3472	0.3299	0.4305	0.3304
H_{13}	0.4457	0.2342	0.2763	0.4693	0.4838	0.3195	0.3548	0.4930	0.3625	0.3442	0.4619	0.3097	1.0000	0.4138	0.5097	0.3582	0.4136	0.4670
H_{14}	0.3925	0.2699	0.2438	0.4546	0.4741	0.3118	0.3360	0.5431	0.4046	0.4710	0.4197	0.3986	0.4138	1.0000	0.4659	0.2679	0.4736	0.3953
H_{15}	0.3201	0.2528	0.2187	0.6215	0.4565	0.3563	0.3391	0.4293	0.3000	0.2855	0.5868	0.3472	0.5097	0.4659	1.0000	0.2437	0.4258	0.5213
H_{16}	0.3631	0.2876	0.2327	0.2176	0.2901	0.3003	0.2312	0.3335	0.2225	0.2982	0.2718	0.3299	0.3582	0.2679	0.2437	1.0000	0.2875	0.2540
H_{17}	0.3058	0.3062	0.1986	0.4162	0.4866	0.3432	0.3625	0.3819	0.3513	0.4911	0.4692	0.4305	0.4136	0.4736	0.4258	0.2875	1.0000	0.3582
H_{18}	0.3523	0.1766	0.1920	0.6266	0.4062	0.2835	0.2565	0.3793	0.2873	0.2311	0.3708	0.3304	0.4670	0.3953	0.5213	0.2540	0.3582	1.0000
$\mu(\kappa_{B\,H_i})$	0.3393	0.2688	0.2275	0.3920	0.3976	0.3162	0.3114	0.3915	0.3067	0.3326	0.3776	0.3533	0.3951	0.3962	0.3930	0.2818	0.3825	0.3464

Table A.2
Agreement between human subjects on the image partitioning task, as measured by the κ_B statistic.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}
X_1	1.0000	-0.0362	-0.0094	-0.5192	-0.7205	-0.6900	-0.2782	-0.1544	-0.3724	-0.4121	-0.3760	-0.3262	-0.2892	-0.2468	-0.1960	-0.3760	-0.2855	-0.3077
X_2	-0.0362	1.0000	0.9511	-0.6651	-0.7704	-0.7251	-0.4832	-0.4499	-0.4425	-0.6245	-0.5626	-0.4961	-0.2449	-0.1526	-0.0925	-0.5534	-0.6032	-0.6623
X_3	-0.0094	0.9511	1.0000	-0.6790	-0.7805	-0.7390	-0.4730	-0.4619	-0.4601	-0.6531	-0.5709	-0.5137	-0.2033	-0.1258	-0.0510	-0.5838	-0.6097	-0.6540
X_4	-0.5192	-0.6651	-0.6790	1.0000	0.6565	0.5964	-0.5044	-0.3253	-0.4213	-0.3114	-0.3142	-0.2495	-0.4656	-0.4601	-0.4573	-0.3400	-0.1904	-0.2754
X_5	-0.7205	-0.7704	-0.7805	0.6565	1.0000	0.8458	-0.6374	-0.4933	-0.5432	-0.4610	-0.4472	-0.4416	-0.6134	-0.6078	-0.6106	-0.5432	-0.4121	-0.4952
X_6	-0.6900	-0.7251	-0.7390	0.5964	0.8458	1.0000	-0.6328	-0.4647	-0.5035	-0.4324	-0.4573	-0.4388	-0.5755	-0.5663	-0.5691	-0.4980	-0.3908	-0.4518
X_7	-0.2782	-0.4832	-0.4730	-0.5044	-0.6374	-0.6328	1.0000	-0.1858	-0.3059	-0.2440	-0.2523	-0.2209	-0.3816	-0.3446	-0.3437	-0.3760	-0.3206	-0.3724
X_8	-0.1544	-0.4499	-0.4619	-0.3253	-0.4933	-0.4647	-0.1858	1.0000	0.2002	-0.0427	-0.1119	0.0488	-0.0380	-0.0380	-0.0500	-0.0990	-0.1415	-0.2301
X_9	-0.3724	-0.4425	-0.4601	-0.4213	-0.5432	-0.5035	-0.3059	0.2002	1.0000	-0.0777	-0.1322	0.0008	-0.0768	-0.0620	-0.0630	-0.0805	-0.2449	-0.3465
X_{10}	-0.4121	-0.6245	-0.6531	-0.3114	-0.4610	-0.4324	-0.2440	-0.0427	-0.0777	1.0000	0.0497	0.1790	-0.2569	-0.2421	-0.3040	-0.0186	0.0405	-0.0907
X_{11}	-0.3760	-0.5626	-0.5709	-0.3142	-0.4472	-0.4573	-0.2523	-0.1119	-0.1322	0.0497	1.0000	0.3702	-0.3188	-0.3151	-0.3289	-0.2560	-0.1821	-0.1175
X_{12}	-0.3262	-0.4961	-0.5137	-0.2495	-0.4416	-0.4388	-0.2209	0.0488	0.0008	0.1790	0.3702	1.0000	-0.1378	-0.1415	-0.1461	0.0137	0.0081	0.0100
X_{13}	-0.2892	-0.2449	-0.2033	-0.4656	-0.6134	-0.5755	-0.3816	-0.0380	-0.0768	-0.2569	-0.3188	-0.1378	1.0000	0.7433	0.6999	-0.2597	-0.2264	-0.3262
X_{14}	-0.2468	-0.1526	-0.1258	-0.4601	-0.6078	-0.5663	-0.3446	-0.0380	-0.0620	-0.2421	-0.3151	-0.1415	0.7433	1.0000	0.6869	-0.2763	-0.2375	-0.3262
X_{15}	-0.1960	-0.0925	-0.0510	-0.4573	-0.6106	-0.5691	-0.3437	-0.0500	-0.0630	-0.3040	-0.3289	-0.1461	0.6999	0.6869	1.0000	-0.2717	-0.2698	-0.3603
X_{16}	-0.3760	-0.5534	-0.5838	-0.3400	-0.5432	-0.4980	-0.3760	-0.0990	-0.0805	-0.0186	-0.2560	0.0137	-0.2597	-0.2763	-0.2717	1.0000	0.3258	0.2390
X_{17}	-0.2855	-0.6032	-0.6097	-0.1904	-0.4121	-0.3908	-0.3206	-0.1415	-0.2449	0.0405	-0.1821	0.0081	-0.2264	-0.2375	-0.2698	0.3258	1.0000	0.4791
X_{18}	-0.3077	-0.6623	-0.6540	-0.2754	-0.4952	-0.4518	-0.3724	-0.2301	-0.3465	-0.0907	-0.1175	0.0100	-0.3262	-0.3262	-0.3603	0.2390	0.4791	1.0000
$\mu(\kappa_{X_i})$	-0.3292	-0.3890	-0.3892	-0.2897	-0.4162	-0.3937	-0.3739	-0.1787	-0.2313	-0.2295	-0.2543	-0.1460	-0.1748	-0.1595	-0.1604	-0.2326	-0.1918	-0.2522

Table A.3
Agreement between computer partitionings of the image set, using a variety of factor analysis techniques, number of retained factors, and images features, as measured by the κ statistic.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}
X_1	1.0000	0.4303	0.4508	0.1211	0.0685	0.0675	0.2684	0.2703	0.1564	0.0864	0.1634	0.1327	0.1676	0.1931	0.2476	0.1399	0.1864	0.2071
X_2	0.4303	1.0000	0.9743	0.0812	0.0745	0.0841	0.1886	0.1408	0.1641	0.0195	0.1024	0.0892	0.2498	0.3042	0.3551	0.0878	0.0491	0.0472
X_3	0.4508	0.9743	1.0000	0.0834	0.0766	0.0848	0.2024	0.1462	0.1651	0.0176	0.1092	0.0932	0.2860	0.3308	0.3884	0.0828	0.0591	0.0640
X_4	0.1211	0.0812	0.0834	1.0000	0.8134	0.7765	0.1351	0.1571	0.1214	0.1459	0.1966	0.1771	0.0474	0.0487	0.0776	0.1574	0.2419	0.2225
X_5	0.0685	0.0745	0.0766	0.8134	1.0000	0.9192	0.1176	0.1337	0.1230	0.1379	0.1842	0.1427	0.0494	0.0511	0.0689	0.1121	0.1801	0.1588
X_6	0.0675	0.0841	0.0848	0.7765	0.9192	1.0000	0.1037	0.1279	0.1250	0.1310	0.1594	0.1194	0.0457	0.0495	0.0694	0.1163	0.1711	0.1644
X_7	0.2684	0.1886	0.2024	0.1351	0.1176	0.1037	1.0000	0.2569	0.2036	0.2026	0.2444	0.2092	0.1162	0.1379	0.1618	0.1471	0.1715	0.1741
X_8	0.2703	0.1408	0.1462	0.1571	0.1337	0.1279	0.2569	1.0000	0.4453	0.2207	0.2409	0.2775	0.2264	0.2238	0.2468	0.2191	0.1741	0.1635
X_9	0.1564	0.1641	0.1651	0.1214	0.1230	0.1250	0.2036	0.4453	1.0000	0.2292	0.2537	0.2751	0.2320	0.2402	0.2674	0.2616	0.1353	0.1156
X_{10}	0.0864	0.0195	0.0176	0.1459	0.1379	0.1310	0.2026	0.2207	0.2292	1.0000	0.3319	0.3518	0.0281	0.0359	0.0327	0.2521	0.2815	0.2365
X_{11}	0.1634	0.1024	0.1092	0.1966	0.1842	0.1594	0.2444	0.2409	0.2537	0.3319	1.0000	0.5511	0.0752	0.0751	0.0982	0.1546	0.1918	0.2759
X_{12}	0.1327	0.0892	0.0932	0.1771	0.1427	0.1194	0.2092	0.2775	0.2751	0.3518	0.5511	1.0000	0.1048	0.0984	0.1364	0.2647	0.2453	0.2975
X_{13}	0.1676	0.2498	0.2860	0.0474	0.0494	0.0457	0.1162	0.2264	0.2320	0.0281	0.0752	0.1048	1.0000	0.8014	0.7780	0.0778	0.0844	0.0740
X_{14}	0.1931	0.3042	0.3308	0.0487	0.0511	0.0495	0.1379	0.2238	0.2402	0.0359	0.0751	0.0984	0.8014	1.0000	0.7677	0.0626	0.0730	0.0713
X_{15}	0.2476	0.3551	0.3884	0.0776	0.0689	0.0694	0.1618	0.2468	0.2674	0.0327	0.0982	0.1364	0.7780	0.7677	1.0000	0.1025	0.0877	0.0805
X_{16}	0.1399	0.0878	0.0828	0.1574	0.1121	0.1163	0.1471	0.2191	0.2616	0.2521	0.1546	0.2647	0.0778	0.0626	0.1025	1.0000	0.5202	0.4897
X_{17}	0.1864	0.0491	0.0591	0.2419	0.1801	0.1711	0.1715	0.1741	0.1353	0.2815	0.1918	0.2453	0.0844	0.0730	0.0877	0.5202	1.0000	0.6453
X_{18}	0.2071	0.0472	0.0640	0.2225	0.1588	0.1644	0.1741	0.1635	0.1156	0.2365	0.2759	0.2975	0.0740	0.0713	0.0805	0.4897	0.6453	1.0000
$\mu(\kappa_{B\,X_i})$	0.1975	0.2025	0.2126	0.2120	0.2007	0.1950	0.1789	0.2159	0.2067	0.1613	0.2005	0.2098	0.2026	0.2097	0.2333	0.1911	0.2058	0.2052

Table A.4
Agreement between computer partitionings of the image set, using a variety of factor analysis techniques, number of retained factors, and images features, as measured by the κ_B statistic.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}	X_{18}	$\mu(\kappa_{B\,H_i})$
H_1	0.1696	0.1581	0.1570	0.0862	0.0622	0.0617	0.1221	0.1195	0.1517	0.0772	0.1527	0.1303	0.1635	0.1316	0.1547	0.0859	0.0903	0.0709	0.1192
H_2	0.1021	0.1282	0.1275	0.0739	0.0620	0.0581	0.1171	0.1258	0.1137	0.0871	0.0589	0.0717	0.1001	0.1009	0.0983	0.1301	0.0810	0.0553	0.0940
H_3	0.1333	0.0970	0.0977	0.1229	0.1199	0.1206	0.0876	0.1576	0.1623	0.0759	0.1031	0.1013	0.0945	0.0982	0.0950	0.0985	0.0993	0.0913	0.1087
H_4	0.0893	0.0735	0.0705	0.0629	0.0541	0.0445	0.1554	0.1021	0.1435	0.0930	0.0971	0.0777	0.1107	0.0840	0.0906	0.1079	0.0698	0.0601	0.0882
H_5	0.1129	0.0937	0.0916	0.0832	0.0574	0.0436	0.1674	0.1420	0.1906	0.1163	0.1163	0.1443	0.1279	0.1151	0.1140	0.1249	0.0862	0.0840	0.1117
H_6	0.1520	0.1346	0.1361	0.0761	0.0494	0.0543	0.1556	0.1327	0.1516	0.0988	0.1177	0.1262	0.1296	0.1200	0.1079	0.1214	0.1022	0.0883	0.1141
H_7	0.0882	0.0750	0.0686	0.0350	0.0268	0.0250	0.0817	0.1257	0.1304	0.0757	0.0981	0.1012	0.1009	0.0941	0.0879	0.0777	0.0732	0.0781	0.0802
H_8	0.1825	0.1252	0.1312	0.1124	0.0831	0.0819	0.1911	0.1254	0.1319	0.0914	0.1539	0.1412	0.1020	0.0719	0.0866	0.0802	0.0912	0.1063	0.1161
H_9	0.1061	0.0683	0.0705	0.0763	0.0510	0.0516	0.1034	0.1419	0.1250	0.1082	0.0819	0.0844	0.0866	0.0918	0.0993	0.1204	0.0853	0.0741	0.0903
H_{10}	0.1234	0.0781	0.0766	0.0693	0.0435	0.0389	0.1241	0.1054	0.1055	0.0814	0.1067	0.0930	0.1089	0.1003	0.1083	0.0790	0.1065	0.1014	0.0917
H_{11}	0.1288	0.1268	0.1236	0.0841	0.0552	0.0630	0.1445	0.1311	0.1424	0.1218	0.1154	0.1181	0.1080	0.1074	0.1205	0.1252	0.0982	0.0771	0.1106
H_{12}	0.1322	0.1693	0.1688	0.0840	0.0736	0.0650	0.1611	0.1383	0.1721	0.0838	0.1233	0.1183	0.1272	0.1208	0.1193	0.0962	0.0871	0.0692	0.1172
H_{13}	0.1472	0.1142	0.1145	0.1102	0.0913	0.0812	0.1519	0.1480	0.1662	0.1280	0.1559	0.1633	0.1623	0.1327	0.1429	0.1251	0.0914	0.1005	0.1293
H_{14}	0.1788	0.1307	0.1294	0.0879	0.0635	0.0574	0.1211	0.1237	0.1397	0.0948	0.1232	0.1042	0.1000	0.0820	0.1133	0.0986	0.1068	0.1127	0.1093
H_{15}	0.0986	0.0830	0.0754	0.0696	0.0483	0.0485	0.1663	0.1382	0.1401	0.1287	0.1157	0.1076	0.1097	0.0967	0.0967	0.1208	0.1136	0.0671	0.1014
H_{16}	0.1295	0.1435	0.1467	0.0713	0.0439	0.0390	0.1066	0.1603	0.1807	0.0851	0.1125	0.1286	0.2312	0.1761	0.2072	0.1328	0.1151	0.1086	0.1288
H_{17}	0.1493	0.0951	0.0990	0.0778	0.0589	0.0490	0.1349	0.1480	0.1460	0.0966	0.1146	0.1069	0.1317	0.1325	0.1300	0.1063	0.0913	0.0764	0.1080
H_{18}	0.1104	0.0994	0.0994	0.0619	0.0519	0.0484	0.1398	0.1409	0.1652	0.1024	0.1335	0.0786	0.1381	0.1065	0.1342	0.0975	0.0558	0.0623	0.1015
$\mu(\kappa_{B\,X_i})$	0.1287	0.1108	0.1102	0.0803	0.0609	0.0573	0.1351	0.1337	0.1477	0.0970	0.1156	0.1109	0.1241	0.1090	0.1170	0.1071	0.0914	0.0824	

Table A.5
Agreement between computer (X_i) and human (S_i) partitionings of the image set, as measured by the κ_B statistic.