Mimicking human texture classification

Eva M. van Rikxoort[a], Egon L. van den Broek[a], and Theo E. Schouten[b]

[a]Nijmegen Institute for Cognition and Information,
Radboud University Nijmegen,
P.O. Box 9104, 6500 HE Nijmegen, The Netherlands
rikxoort@sci.ru.nl, e.vandenbroek@nici.ru.nl
http://eidetic.ai.ru.nl/egon/

[b]Nijmegen Institute for Computing and Information Science,
Radboud University Nijmegen,
P.O. Box 9010, 6500 GL Nijmegen, The Netherlands
T.Schouten@cs.ru.nl
http://www.cs.ru.nl/~ths/

ABSTRACT

In an attempt to mimic human (colorful) texture classification by a clustering algorithm three lines of research have been encountered, in which as test set 180 texture images (both their color and gray-scale equivalent) were drawn from the OuTex and VisTex databases. First, a k-means algorithm was applied with three feature vectors, based on color/gray values, four texture features, and their combination. Second, 18 participants clustered the images using a newly developed card sorting program. The mutual agreement between the participants was 57% and 56% and between the algorithm and the participants it was 47% and 45%, for respectively color and gray-scale texture images. Third, in a benchmark, 30 participants judged the algorithms' clusters with gray-scale textures as more homogeneous then those with colored textures. However, a high interpersonal variability was present for both the color and the gray-scale clusters. So, despite the promising results, it is questionable whether average human texture classification can be mimicked (if it exists at all).

Keywords: Human texture perception, mimic, k-means, color, texture, card sorting, clustering

1. INTRODUCTION

Most computer vision (CV) and content-based image retrieval (CBIR) systems[1–5] rely on the analysis of features such as color, texture, shape, and spatial characteristics. Some of these CV and CBIR systems are partly founded on principles known from human perception. However, these systems are seldomly validated with experiments where humans judge their artificial counterparts. The current paper discusses the process of such a validation for the artificial analysis of texture (i.e., mimicking human texture classification). In addition, the influence of color on texture classification is a topic of this research.

As feature for the human visual system, texture reveals scene depth and surface orientation; moreover, it describes properties like the smoothness, coarseness, and regularity of a region (c.f. Rao and Lohse[6], Battiato, Gallo, and Nicotra[7], and Van Rikxoort and Van den Broek[8]). Texture is efficiently encoded by the human visual system; as Bergen and Adelson[9] stated: "... simple filtering processes operating directly on the image intensities can sometimes have surprisingly good explanatory power." Inspired by human texture processing, artificial texture analysis techniques describe similar properties as human perception does. However, direct comparisons between human and artificial texture processing are seldomly made.

In 2000, Payne, Hepplewhite, and Stonham[10] presented research toward mimicking human texture classification. Given a target image, they asked 30 humans to classify textures. Next, they compared these classifications with the classifications done by several texture analysis techniques. They concluded that, where the human

Send correspondence to Egon L. van den Broek, E-mail: e.vandenbroek@nici.ru.nl

Human Vision and Electronic Imaging X, edited by Bernice E. Rogowitz,
Thrasyvoulos N. Pappas, Scott J. Daly, Proc. of SPIE-IS&T Electronic Imaging,
SPIE Vol. 5666 © 2005 SPIE and IS&T · 0277-786X/05/$15

visual system works well for many different textures, most texture analysis techniques do not. For only 20%-25% of the textures, a match was found between artificial and human classification.

The research of Payne et al.[10], as most research in CV and CBIR, concerned gray-scale texture analysis. This despite that most image material is in color. As Palm[11] already denoted: "The integration of color and texture is still exceptional." From a scientific point of view, one can argue that since neither texture nor color is fully understood, the influence on each other is simply too unpredictable to do research in, at least outside a controlled experimental environment.

Color on its own, already is a complex phenomenon, as is texture. The perception of color is influenced by both environmental issues (e.g., position of the light source and properties of material) and internal processes present in the observer (e.g., color constancy). However, concerning color classification or categorization, evidence is present for the existence of 11 color categories[12–16] (i.e., black, white, red, green, yellow, blue, brown, purple, pink, orange, and gray), used by human memory. Recently, this concept was embraced and utilized for the development of color-based image retrieval[5].

Our approach to mimicking human (colorful) texture classification is different from the approach used by Payne et al.[10] First, we let a k-means algorithm cluster the whole dataset using different feature vectors (Section 3). Next, in Section 4, we let humans cluster the whole dataset and in Section 5, we determine to which extend our k-means clusterings mimic the human clustering. As a follow up study, we let humans judge the clusters generated by the artificial clustering techniques (Section 6). We conclude with a discussion in Section 7.

2. EXPERIMENTAL SETUP

In this section, general specifications are provided, which hold for all three experiments (i.e. automatic clustering, human clustering, and humans judging the automatic clustering). As data, a collection of 180 colorful texture images were drawn from the OuTex and VisTex databases. Two criteria were used when selecting the images: (i) there had to be images from at least fifteen different categories and (ii), when a class was extremely large compared to the other classes, only a subset of the class is used. These criteria make sure the task of clustering the images is not trivial. Moreover, the images were resized in order to fit on one screen. This was needed to facilitate an optimal and pleasant execution of the experiment. Figure 1 provides an overview of all the 180 images.

In both the first and the second line of research, two experiments were ran: one with the original color images and one with gray versions of the same images. To obtain the latter, the set of 180 images was converted to gray-scale (I) images (i.e., $I = (R + G + B)/3$). Now, two identical sets of images were present, except for presence versus absence of color information.

Clustering of images can be seen as sorting the images in a number of categories or stacks. So, the clustering of texture images can be treated as a card sorting task[17]. In such a task, the participant is asked to sort cards (e.g., images) and put them on separate stacks. As a consequence, only the top image on each stack is visible. So, participants have to memorize a representation of each of the stacks they defined. However, during the task the number of images on the stacks will increase and the content of the stack will change. Therefore, also the representation of the stacks needs to be updated, for which the human visual Short Term Memory (vSTM)[18] has to be taken into account.

Human vSTM can contain four[19] to fourteen[20] items. The number of clusters made by humans needs to be within this range. To be able to compare the clusters of textures made by the participants, they all had to define the same number of clusters. Moreover, the automatic clustering also had to result in the same number of clusters in order to be able to compare it with its human counter parts.

To determine this number of clusters, we asked five experts to cluster the images in an arbitrary number of clusters, with an upper limit of fourteen. The mean number of clusters produced by the experts is taken as the number of clusters to be produced. The experts determined the optimal number of clusters for this dataset, on both the gray-value and colorful images, to be six. Please note that, on the one hand, this is on the safe side of this interval but, on the other hand, the images were taken from fifteen different categories, not from six.

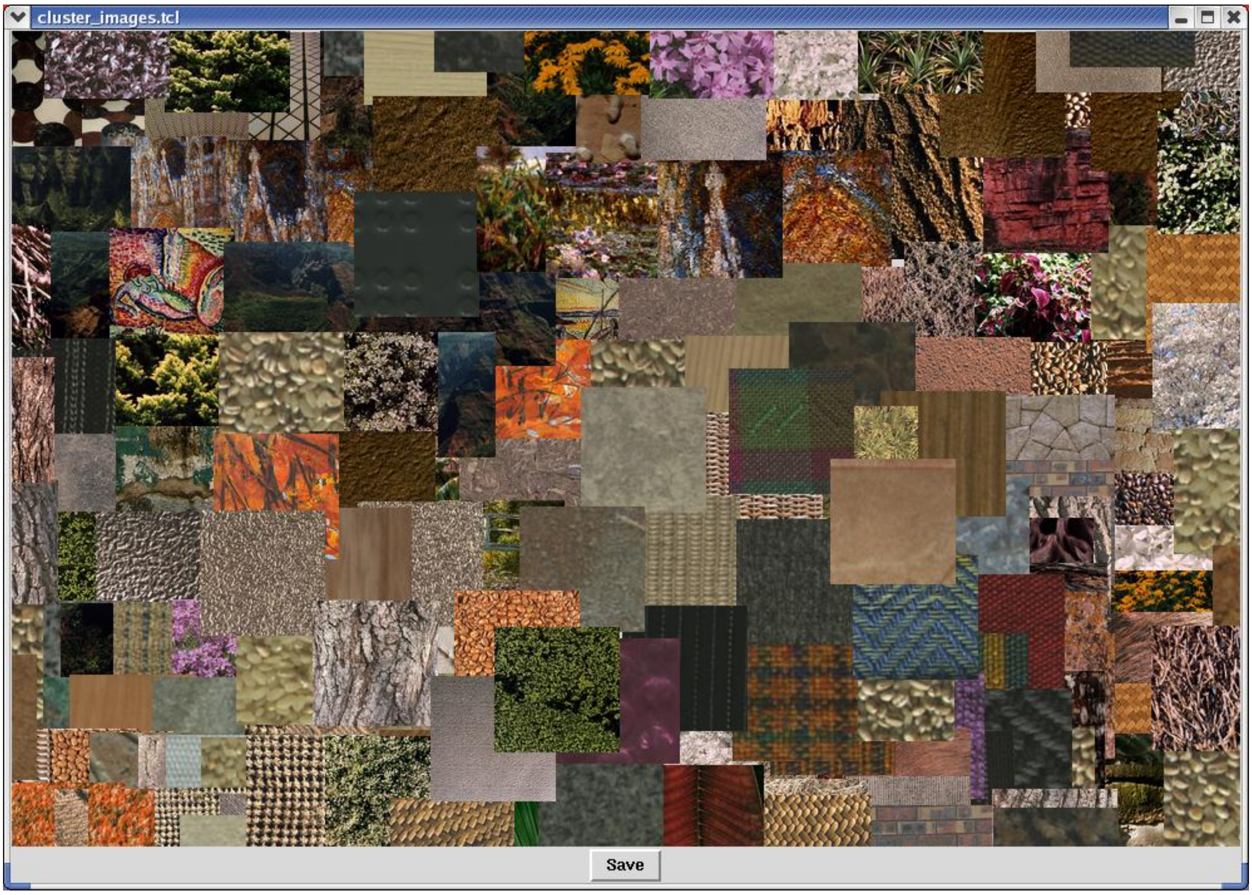


Figure 1. An overview of all 180 images (the color version) used in the clustering experiments with both human participants and the automatic classifier.

3. AUTOMATIC TEXTURE CLUSTERING

Automatic texture clustering is done by a three step approach for each of our two sets of images: (1) defining a suitable feature space, (2) calculate the feature vector of each image, such that each image is represented by a point in the feature space, (3) find groups or clusters of points in the feature space.

Many approaches have been developed for clustering points in feature space; see Mitchel[21] and Berkhin[22] for recent surveys. These approaches can be divided in two groups from the perspective whether or not additional information on data points is available. Supervised approaches need, at least for a representative sample of the data points, information to which cluster each data point belongs. In our case this would mean dividing the data set provided by the human clustering into two or three parts: a part used for training a supervised method, a part for evaluating the parameters used during training (this is often not done as the available supervised data set is usually small), and a part for evaluating the final result of the clustering. In our case the data set is too small to allow splitting it into parts.

Unsupervised methods do not need labeling of the data points. But usually they require the number of clusters as additional input. Either they use it as a fixed a priori number needed to start the clustering process, or they use it as a termination condition, otherwise they would continue until each data point is its own cluster. In our case, the number of intrinsic clusters, was determined by experts (Section 2), who determined the output to be six clusters. This enables us to compare the automatic clustering to the human clustering.

Since we did not have any information on the distribution of the points in our feature space, we evaluated two general applicable and often used methods: hierarchical clustering and k-means clustering. Evaluation of these two methods on a early available subset of our data did not show a preference for one of the two; the produced results were comparable. For this paper, we chose to use the k-means method as it has somewhat more possibilities to tune certain parameters.

In this research, three different feature vectors are used for the k-means algorithm. In a previous research[23, 24], we determined the optimal configurations for both colorful and gray-scale texture classification. The optimal configuration for colorful texture analysis turned out to be our new parallel-sequential approach, using four texture features (i.e., entropy, inverse difference moment, cluster prominence, and Haralick's correlation), from the color correlogram[25] based on the 11 color categories[12–16]. For gray-scale texture analysis, the parallel approach performed best, in which the four texture features from the co-occurrence matrix[26–28] are combined with a gray-scale histogram from the HSV color space quantized in 32 bins. For more detailed information, we refer to Van den Broek and Van Rikxoort[23, 24].

In this experiment, for both color and gray-scale, k-means clustering was applied using three different feature vector configurations consisting of: (i) color or gray-scale information (i.e., the histogram), (ii) textural information (i.e., the four texture features), and (iii) both color and texture information (i.e., the histogram and the four texture features).

For each of the six vectors used in the k-means clustering, six clusters of images resulted. In Table 1 the size of each of the clusters is shown.

Table 1. The size of the six clusters constructed by the k-means algorithm for the different feature vectors for both color and gray-scale.

Feature vector	**Color**						**Gray-sclae**					
color/gray-scale features	29	29	30	25	29	38	25	33	13	18	38	53
texture features	17	18	68	13	15	49	3	19	66	20	43	29
combined features	42	25	24	25	28	36	15	14	49	28	32	42

4. HUMAN TEXTURE CLUSTERING

4.1. Method

Eighteen subjects with normal or corrected-to-normal vision and no color deficiencies participated. They all participated on a voluntary basis. Their age ranged from 16 to 60. Half of them were male and half of them were female. All participants were naive with respect to the goal of the research. In addition, neither of them was specialized in color or texture perception.

The experiments were executed on multiple PCs. In all cases the screen of the PC was set on a resolution of 1024 × 768 pixels. Moreover, we assured that the experiment was conducted in an average office lighting. We chose for this loosely controlled setup of apparatus, since it represented an average office situation and our opinion is that good algorithms mimicking human perception should be generally applicable and robust enough to handle images, which are taken and viewed under various circumstances. To put it in a nutshell, we consider the world as our experimental environment.

Two experiments were conducted. They differed only with respect to the stimuli; i.e., the texture images (see Section 2 and Figure 1). In one of the experiments color images were presented; in the other experiment their gray equivalents were presented (see also Section 2). In order to control for possible order effects, half of the participants executed the experiments in the one order and the other half in the other order.

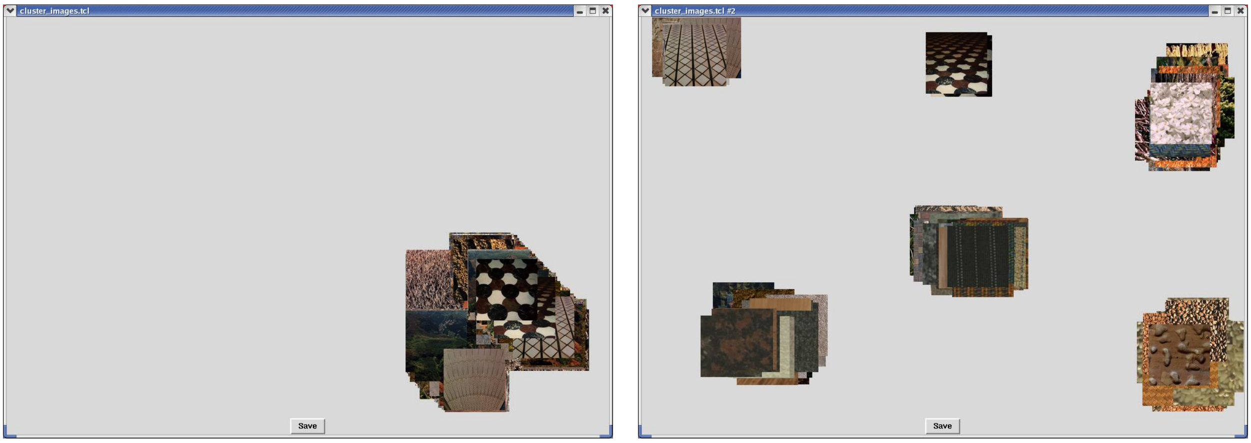
As discussed in Section 2, clustering of images can be represented as a card sorting task. However, in order to control and automate the sorting task as much as possible, a Tcl/Tk program was used that fully operationalized the desktop metaphor. A canvas (i.e., window) was presented on a computer screen in which the images can be moved and stacked on each other, just like on a regular desktop[29, 30].

At the start of the experiments, the images are shown as a pile on the canvas. To tackle possible effects in sorting due to the order of presentation of the images, the images were placed in random order on the pile. So, at the start of the experiment, the canvas presented one pile of 180 randomly sorted images, as is shown in Figure 2a.

The participants were able to drag the images by way of a mouse. They were allowed to drag the images all over the screen and drop them on any position wanted. During the experiment, all images were free to be positioned otherwise and, so, it was possible to change, merge, or divide already defined stacks. The only restriction was that the six resulting stacks were placed clearly separately from each other in order to tackle possible overlap between stacks. An example of such a final cluster is provided in Figure 2b.

The participants were not instructed what features to use for the classification. This loose instruction guaranteed an unbiased human texture classification. The latter was of the utmost importance since we wanted to mimic human texture classification and were not primarily interested in the underlying (un)conscious decision making process.

After a definite choice of clusters was determined, the result was saved (by pressing the save button). For each image, its coordinates as well as its name were saved. Hence, the stacks could be reproduced, visualized, and analyzed easily.



(a)

(b)

Figure 2. (a) The start condition of the experiment: one pile of 180 images. (b) An example of a final result of an experiment: six clusters of images.

4.2. Data analysis

For both experiments, the same data analysis was applied. In this section, the data analysis is described; in the next section, the results are presented.

For each of the 153 $(18!/(16! \cdot 2!))$ unique pairs of participants (p_i, p_j) a consensus matrix $(M_{(p_i,p_j)})$ of size 6×6 was determined, which contains for each pair of clusters, the number of matching images. Non-unique pairs of clusters were chosen since one cluster of participant i can encapsulate the images assigned to two separate clusters by a participant j and vice versa. From the set of confusion matrices, two data were derived: (i) the average consensus on the clustering between participants and (ii) the most prototypical set of clusters, in other words the most prototypical participant.

The average consensus in the clustering between participants was determined as follows: For each pair of participants (p_i, p_j), the consensus $C_{(p_i,p_j)}$ is determined by summing the highest value of each of the six rows of the consensus matrix $M_{(p_i,p_j)}$. So, $C_{(p_i,p_j)} = \sum_{i=1}^{6} \; max\{row_i\}$. Now, the overall consensus can be determined by: $\sum_{p_ip_j} C_{(p_ip_j)}/153$.

Of particular interest is the most prototypical set of clusters since it describes the most prototypical human (clustering); i.e., the highest average consensus to all other participants. The average consensus A of participant p_i is defined by: $A_{p_i} = \sum_{j=1}^{18} C_{(p_i,p_j)}$. Subsequently, the most prototypical participant, or the most prototypical set of clusters, can be defined as: $C_{ave} = max\{A_{p_i}\}$. To be able to describe these six clusters, we determined for all images the number of times they were assigned to one of the prototypical clusters.

4.3. Results

4.3.1. Colorful texture clustering

The average consensus between the participants with respect to colorful textures was 57%. The consensus between the participants ranged from 39% to 87%. The consensus matrix describing the consensus between all pairs of participants (see in Table 2 the numbers in a normal font) illustrates the variance present between the participants, in the clusterings.

Table 2. The consensus on the clustering between the 18 participants (p). The numbers in a normal font denote the colorful images; the numbers in the italic font denote the gray-scale images.

	p01	p02	p03	p04	p05	p06	p07	p08	p09	p10	p11	p12	p13	p14	p15	p16	p17	p18
p01	–	66	60	68	54	39	50	63	46	79	50	53	64	50	51	54	69	49
p02	*50*	–	67	74	50	39	50	56	51	78	48	45	58	44	49	56	46	60
p03	*63*	*55*	–	60	60	40	48	60	45	81	50	50	61	47	51	56	58	54
p04	*54*	*53*	*56*	–	50	38	49	70	54	87	53	55	67	44	57	60	63	61
p05	*48*	*51*	*50*	*42*	–	36	40	55	40	68	45	44	49	46	55	52	51	54
p06	*46*	*37*	*41*	*40*	*40*	–	38	49	45	70	42	41	49	39	46	42	41	37
p07	*54*	*58*	*78*	*53*	*50*	*43*	–	63	55	87	52	50	59	54	50	55	52	53
p08	*58*	*59*	*71*	*59*	*53*	*46*	*68*	–	56	83	54	54	56	49	50	53	62	50
p09	*52*	*51*	*65*	*47*	*46*	*36*	*60*	*53*	–	78	49	53	59	42	52	51	49	59
p10	*46*	*46*	*56*	*43*	*49*	*42*	*53*	*45*	*47*	–	51	51	54	41	54	54	51	51
p11	*68*	*55*	*70*	*63*	*50*	*54*	*63*	*66*	*64*	*64*	–	55	63	52	54	61	59	61
p12	*52*	*55*	*63*	*61*	*51*	*44*	*57*	*57*	*48*	*46*	*49*	–	55	47	49	50	50	54
p13	*60*	*45*	*56*	*55*	*43*	*41*	*47*	*49*	*45*	*49*	*64*	*58*	–	51	54	59	72	64
p14	*55*	*54*	*66*	*51*	*48*	*44*	*65*	*61*	*45*	*51*	*58*	*56*	*59*	–	50	56	49	59
p15	*47*	*53*	*57*	*47*	*60*	*49*	*53*	*54*	*51*	*49*	*47*	*52*	*50*	*56*	–	47	45	45
p16	*47*	*50*	*55*	*51*	*45*	*40*	*49*	*49*	*41*	*51*	*49*	*56*	*62*	*51*	*51*	–	52	65
p17	*65*	*53*	*60*	*61*	*45*	*40*	*51*	*55*	*41*	*40*	*51*	*61*	*74*	*50*	*44*	*49*	–	48
p18	*58*	*58*	*70*	*57*	*53*	*45*	*59*	*60*	*53*	*56*	*55*	*64*	*64*	*61*	*55*	*60*	*63*	–

In order to establish the most prototypical set of clusters, we determined a set of core images on which at least 45% of the participants agreed. This approach is adapted from Payne et al.[10]. The treshold of 45% was chosen because the clustering is probably a fuzzy one, so with this treshold, images can be assigned to two different clusters and still be a core image. For the colorful textures, this resulted in a set of 88 (out of 180) core images. The overall, average consensus between the participants on the core images was 70%. Next, the most prototypical participant was determined, based on the set of core images. One participant did have an average consensus of 82% with all other participants; hence, the clusters of this participant are labeled as prototypical clusters.

The prototypical clusters are now used to determine the base images for all prototypical clusters. An image is said to be a base image for a particular cluster if it is assigned to the cluster by at least 8 ((18 / 2) - 1) participants. The clusters can be described by respectively, 37, 26, 14, 37, 37, and 45 base images. Moreover, 24 images appeared to be a base image for more then one cluster. The mean frequency of the base images in the clusters is 11.74.

4.3.2. Gray-value texture clustering

The average consensus between the participants with respect to gray-value textures was 56%, ranging from 37% to 78%. In Table 2, the numbers in an italic font, provide the consensus between all pairs of participants. As Table 2 illustrates, a considerable amount of variance is present in the consensus between the participants on the clustering of gray-value textures.

For the determination of the core images, again a threshold of 45% was chosen. This resulted in a set of 95 (out of 180) core images, which is slightly more than with the color textures. In contrast, the average consensus between the participants on the gray-value images was slightly less (65%) than with the color textures. The participant assigned as prototypical, did have an average consensus of 73% to all other participants.

The clusters can be described by respectively, 32, 21, 32, 44, 24, and 46 base images. Moreover, 42 images appeared to be a base image for more then one cluster. The mean frequency of the base images in the clusters is 12.01.

5. AUTOMATIC VERSUS HUMAN TEXTURE CLUSTERING

Since the goal of this research is to mimic human texture classification, we want to compare the automatically generated clusters to the clusters generated by the human participants. The same analysis is applied for the colorful textures and the gray-scale textures. For both the clusters of color and gray-scale images, each of the 54 (18·3) unique pairs of participant - automatic clusterings, a consensus matrix was constructed (see Section 4.2). Two types of similarity were derived from these matrices: (i) the overall consensus between the automatic clusterings and the human clusterings and (ii) the consensus based on the clusters defined by their base images (see Section 4.3).

5.1. Data analysis

The consensus between the automatic and the human clusterings was determined as described in Section 4.2 with (p_i, p_j) being a pair of participant - automatic clustering instead of a pair of participants.

Next to the average consensus, the consensus on the prototypical clustering (as described in Section 4.3) is of interest. For this purpose, we will now define: a binary measure and a weighted measure.

The binary measure of agreement assigns one cluster (c) to each image (I) by means of the frequency of assignment by the participants (see Section 4.3). The cluster with the highest frequency of assignment is assigned to the image (I_c). This clustering is compared to the automatic clusterings for each image (I_a) in a binary way. Let ϕ be the binary value assigned to each image. Then, for each image I, ϕ is 1 when $I_c = I_a$ and ϕ is 0 when $I_c \neq I_a$. The total binary agreement is now defined by $\sum_\phi$. Last, the binary agreement for each cluster x is defined by $\sum_\phi |c = x$. The binary agreement is normalized by dividing it by the number of images.

The weighted measure of agreement weights the agreement on the clustering and is based on the frequencies of assignment to a cluster by humans. The frequencies are divided in four categories, the first category has a frequency of at least15, the second category has a frequency of at least 11, the third category has a frequency of at least 7, and finally the fourth category has a frequency less than 7. Let θ be the weighted measurement value for each image. Then, for each image I, θ is 3 when I_a is in the first category, θ is 2 when I_a is in the second category, θ is 1 when I_a is in the third category, and θ is 0 when I_a is in the last category. The total weighted agreement is now defined by $\sum_\theta$. The weighted agreement for each cluster x is defined by $\sum_\theta |c = x$. The weighted agreement is normalized by dividing it by the total weighted agreement of the most optimal clustering.

The weighted measure is used next to the binary (standard) measure because the human clustering is a fuzzy one and is only defined by the frequencies of assignment. In the binary measure, no use is made of these frequencies.

Table 3. The percentages of correct classification for the colorful images for each cluster and for the whole dataset using the binary measure and the weighted measure.

cluster	binary measure			weighted measure		
	color	**texture**	**combined**	**gray**	**texture**	**combined**
1	35%	22%	32%	50%	50%	50%
2	42%	46%	42%	39%	62%	61%
3	0%*	0%*	0%*	100%*	100%*	100%*
4	22%	16%	22%	58%	73%	69%
5	76%	54%	60%	60%	83%	45%
6	40%	53%	53%	85%	43%	71%
All images	44%	39%	43%	42%	44%	45%

5.2. Results

5.2.1. Colorful textures

For the colorful textures, three configurations (i.e., feature vectors) for k-means clustering were used (see Section 3): (i) the 11 color histogram, (ii) the four texture features, and (iii) a combination of the color and texture features, resulting in a feature vector of length 15.

For each of the three feature vectors, its average consensus with the participants' clusters was determined, as described in Section 4. The average consensus between human and automatic clustering using only color information was 45%, using only texture information it was 46%, and using both color and texture information it was 47%.

In Table 3, the results from the binary and weighted measures of agreement, between human and automatic clustering are given. It is possible that no images are assigned to a particular human cluster because we adapted the same approach for the calculation of the consensus as described in Section 4: non-unique mapping of the clusters. The percentages marked with a * in Table 3 are the result of the fact that no images were assigned to the particular cluster by the specific automatic clustering.

For the binary measure, there are two clusters on which one of the feature vectors had a percentage of more than 50%. For the weighted measure, four clusters present a consensus of more than 50% between human and artificial clusterings (see also Table 3).

5.2.2. Gray-scale textures

For the gray-scale textures, three configurations (i.e., feature vectors) for k-means clustering were used (see Section 3): (i) the 32 bins HSV gray-scale histogram, (ii) the four texture features, and (iii) a combination of the histogram and texture features, resulting in a feature vector of length 36.

For each configuration of automatic clustering, its average consensus with the participants' clusters was determined, as described in Section 4. The average consensus on the automatic clustering using only gray-scale information was 44%, using only texture information it was 45%, and using gray-scale and texture information it was 42%.

In Table 4 the results from the binary and weighted measures of agreement, between human and automatic clustering are given. For the binary measure, there are four clusters on which one of the automatic classifiers had a percentage of more than 50%. For the weighted measure, even five clusters present a consensus of more than 50% between human and artificial clustering.

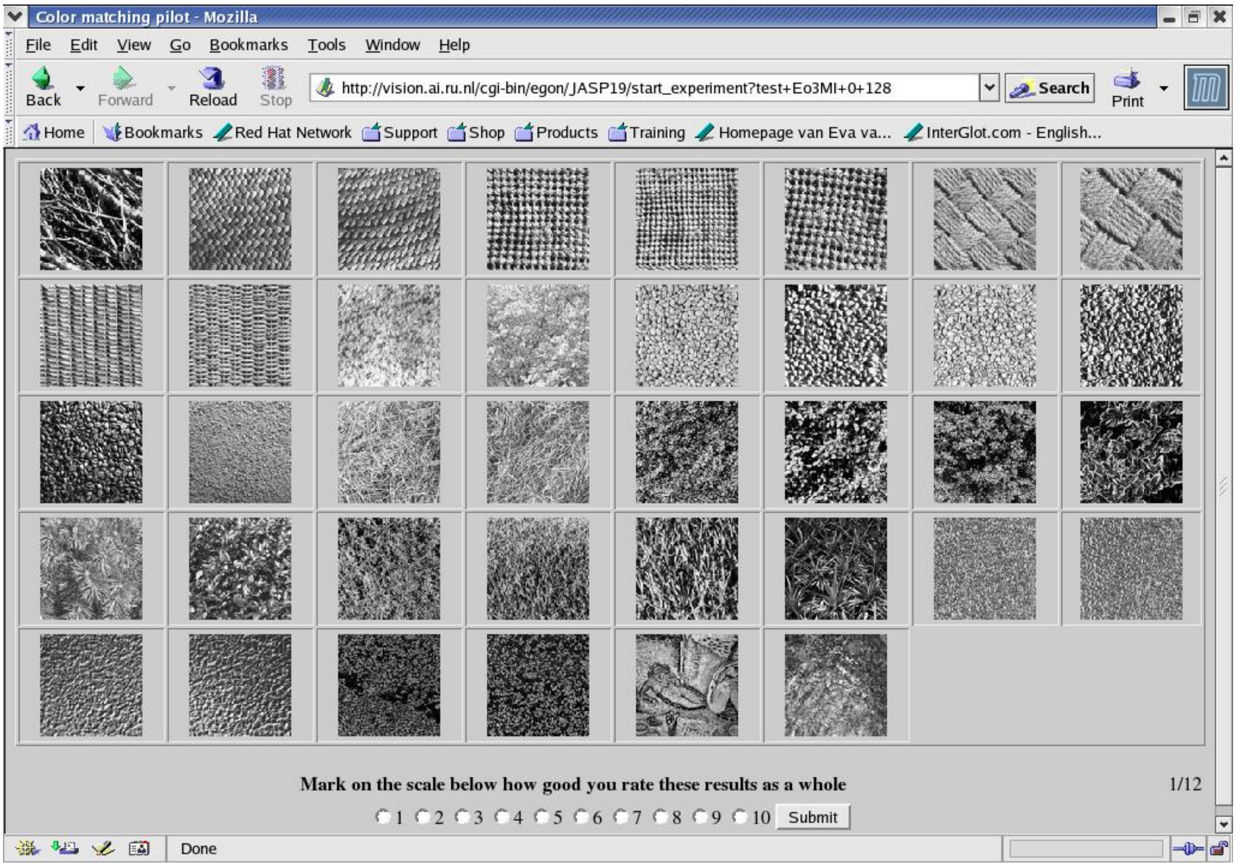
Considering that the clusters to which no images are assigned by the automatic classifier are left out, this means that for the gray-scale images, on all clusters a percentage of more than 50% for the weighted measure was achieved.

Table 4. The percentages of correct classification on the gray-scale images for each cluster and for the whole dataset using the binary measure and the weighted measure.

cluster	binary measure			weighted measure		
	gray	**texture**	**combined**	**gray**	**texture**	**combined**
1	100%	44%	50%	97%	47%	62%
2	52%	0%	62%	100%	70%	59%
3	0%	0%	0%	100*%	100*%	100*%
4	61%	68%	68%	79%	65%	71%
5	88%	0%	83%	100%	100*%	100%
6	0%	7%	0%	100*%	100%	100*%
All images	36%	41%	44%	33%	41%	43%

6. HUMANS JUDGING AUTOMATIC CLUSTERING

As a follow up experiment, humans were asked to judge the clusters generated by the automatic clustering algorithm. For both color and texture, the clusters of the automatic clustering algorithm with the best average performance were chosen; for color the k-means algorithm using color and texture features, and for gray-scale the k-means algorithm using only texture features.



Color matching pilot - Mozilla

http://vision.ai.ru.nl/cgi-bin/egon/JASP19/start_experiment?test+Eo3MI+0+128

Mark on the scale below how good you rate these results as a whole

1/12

○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10 Submit

Done

Figure 3. An example screen from the benchmark used to let users judge the automatic clusters.

For this experiment, a benchmark was used which was introduced by Van den Broek, Kisters, and Vuurpijl[5]. It allowed users to judge each individual cluster for its homogeneity and correctness. The benchmark showed all images of a cluster in one screen, at the bottom of the screen a mark between 1 and 10 can be given for the homogeneity of the cluster shown. All users are presented with 12 screens; 6 containing gray-scale images and 6 containing colorful images. An example screen from the benchmark is shown in Figure 3.

In this experiment, 36 subjects, with normal or corrected-to-normal vision and no color deficiencies participated. Their participation was on a voluntary basis and they were naive with respect to the goal of the research. The age of the participants varied from 18 to 60, half of the participants were male and half of them were female.

The experiment was run on line and can be found on `http://eidetic.ai.ru.nl/egon/HVEI2005-Texture_benchmark`. The participants were instructed to judge the clusters on their homogeneity. They were not informed about the clusters being produced by artificial classifiers.

For both the colorful texture clusters and the gray-scale texture clusters, we determined average rating given for the homogeneity of the results. The average rating for the gray-scale clusters was 6.1, with a standard deviation of 3.1; the average rating for the colorful clusters was 5.2, also with a standard deviation of 3.1. The gray-scale clusters were judged significantly better than the colorful clusters ($p < .0069$). The high standard deviations of the ratings denote a high variation between the participants in judging the clusters.

7. DISCUSSION

In the present research, first a set of 180 texture images were clustered by a k-means clustering algorithm, using three different feature vectors for both color and gray-scale. Next, 18 humans were asked to cluster the set of texture images both in gray-scale and color. Using the clusterings of all participants, a set of base images for each cluster was derived, which describe the clusters. The automatic clusterings were compared to the human clustering using two measures of agreement (i.e., binary and weighted). In addition, the influence of color compared to gray-scale was investigated. Last, a benchmark was executed in which 36 participants judged the automatic clustering results.

For both the colorful textures and the gray-scale textures, little consensus was present between the participants. Although all participants reported more trouble clustering the gray-scale images, the consensus between the participants was almost the same on the colorful textures and the gray-scale textures (57% vs 56%). The low consensus between the participants indicates that the task of clustering the textures selected was not a trivial one, as was our aim in selecting the images (see Section 4).

The overall success in comparing the automatic classifier to the human classifications was the same for the colorful textures and the gray-scale textures (45% - 47% versus 42% - 45%). When inspecting the results for the separate clusters however, more success is shown on the gray-scale clusters. For the gray-scale textures, using the binary measure of agreement, for four clusters more than 50% of the images were classified correct. The weighted measure for the gray-scale images gives a good result on five of the clusters. The mean percentages of correct classification for the clusters, which are matched well, are 76% and 95% for the binary and weighted measure respectively.

For the colorful textures, there are respectively two and four clusters that match well. The mean percentages of correct classification for the clusters which are matched well are 65% and 80% for the binary and weighted measure respectively. So, the match in clustering between humans and the k-means algorithm is more convincing for the gray-scale images than for the colorful images. This effect of overall performance versus cluster-performance is caused by the non-unique mappings we used to determine the consensus between clusters. For gray-scale, there are six instances in which no images are assigned to a particular cluster (see Table 4) which impairs the results over the overall dataset. Moreover, for the clusters to which images are assigned, good to excellent results are obtained. For the colorful images, there are only three instances in which no images are assigned to a particular cluster, where the results for the other clusters are not convincing either.

An inspection of the images itself revealed that the clusters that are mimicked well by the automatic classifiers, show little variation in color/gray-scale and texture. So, all images in a well mimicked cluster, have the same texture properties like randomness, directionality, and coarseness, and show little variation in color/gray-scale. For both gray-scale and color, the cluster to which no images were matched by the automatic classifiers, seem to be a ‘garbage group’, in which the human participants put all images they were unable to label. Such a ‘garbage group’ was mentioned by all participants.

The fact that the gray-scale images are better mimicked by the automatic clustering methods can partly be explained by the trouble humans reported in clustering the gray-scale images. These difficulties in clustering were mainly caused by the fact that on the gray-scale images less semantics were visible. So, on the gray-scale images humans use more pattern and gray-scale based clustering than semantic based clustering. In contrast, on the colorful images, most humans used semantic features for clustering.

For both the colorful textures and the gray-scale textures, three feature vectors were used for the k-means clustering method, utilizing respectively color/gray-scale information, texture information, and their combination. For the colorful textures there is no significant difference between the different feature vectors on the results of mimicking human texture perception. However, for the gray-scale textures, the feature vector using only gray-scale information performs worse than the other two feature vectors.

Next to the fact that human gray-scale clustering is better mimicked by our artificial classifier, the automatic clusters for the gray-scale textures are also significantly better rated by humans than the colorful texture clusters of the automatic classification. In earlier research, we found that using color improves classification performance based on (human) semantic labeling of the dataset[23, 24]. So, a possible conclusion that can be drawn from this result is that humans use more than just semantic when clustering textures.

Although human gray-scale texture clustering was better mimicked by the automatic clustering, the results on colorful texture clustering were also satisfying since the mean correct classification is twice as high as the classification reported by Payne et al.[10]. So, despite the low percentages of consensus between humans and the clustering algorithm, the results should be considered as promising. With that, this research presents a successful first attempt to mimic human colorful texture classification.

ACKNOWLEDGMENTS

The Dutch organization for scientific research (*NWO*) is gratefully acknowledged for funding *Eidetic* (project-number: 634.000.001), a project within the *ToKeN* research line, in which this research was done. Furthermore, we would like to Merijn van Erp for reviewing earlier versions of the data analysis sections and Louis Vuurpijl for the initial version of Tcl/Tk application which we used.

REFERENCES

1. M. Israël, E. L. van den Broek, P. van der Putten, and M. J. den Uyl, "Automating the construction of scene classifiers for content-based video retrieval," in *Proceedings of the Fifth International Workshop on Multimedia Data Mining (MDM/KDD'04)*, L. Khan and V. A. Petrushin, eds., pp. 38–47, (Seattle, WA, USA), 2004.
2. W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Petkovic, P. Yanker, and C. Faloutos, "The QBIC project: Querying images by content using color, texture, and shape," in *Proceedings of Storage and Retrieval for Image and Video Databases*, W. Niblack, ed., **1908**, pp. 173–187, February 1993.
3. A. Pentland, R. Picard, and S. Sclaroff, "Photobook: Tools for content based manipulation of image databases," in *Proceedings of SPIE Storage and Retrieval for Image and Video Databases II, Electronic Imaging: Science and Technology* **2185**, pp. 34–47, (San Jose, CA, USA), 1994.
4. A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(12), pp. 1349–1380, 2000.
5. E. L. van den Broek, P. M. F. Kisters, and L. G. Vuurpijl, "The utilization of human color categorization for content-based image retrieval," in *Proceedings of Human Vision and Electronic Imaging IX*, B. E. Rogowitz and T. N. Pappas, eds., **5292**, pp. 351–362, (San Jose, CA, USA), 2004.
6. A. Ravishankar Rao and G. L. Lohse, "Towards a texture naming system: Identifying relevant dimensions of texture," *Vision Research* **36**(11), pp. 1649–1669, 1996.
7. S. Battiato, G. Gallo, and S. Nicotra, "Perceptive visual texture classification and retrieval," in *Proceedings of the 12th International Conference on Image Analysis and Processing*, M. Ferretti and M. G. Albanesi, eds., pp. 524–529, (Mantova, Italy), September, 17-19 2003.

8. E. M. van Rikxoort and E. L. van den Broek, "Texture analysis." Technical report, NICI, Radboud University Nijmegen; URL: `http://eidetic.ai.ru.nl/egon/publications/pdf/Rikxoort04-Texture_analysis.pdf`, 2004.
9. J. R. Bergen and E. H. Adelson, "Early vision and texture perception," *Nature* **333**(6171), pp. 363–364, 1988.
10. J. S. Payne, L. Hepplewhite, and T. J. Stoneham, "Applying perceptually-based metrics to textural image retrieval methods," in *Proceedings of Human Vision and Electronic Imaging V*, B. E. Rogowitz and T. N. Pappas, eds., **3959**, pp. 423–433, (San Jose, CA, USA), 2000.
11. C. Palm, "Color texture classification by integrative co-occurrence matrices," *Pattern Recognition* **37**(5), pp. 965–976, 2004.
12. B. Berlin and P. Kay, *Basic color terms: Their universals and evolution*, Berkeley: University of California Press, 1969.
13. R. M. Boynton and C. X. Olson, "Locating basic colors in the osa space," *Color Research & Application* **12**, pp. 107–123, 1987.
14. E. L. van den Broek, M. A. Hendriks, M. J. H. Puts, and L. G. Vuurpijl, "Modeling human color categorization: Color discrimination and color memory," in *Proceedings of the Fifteenth Belgium-Netherlands Conference on Artificial Intelligence*, T. Heskes, P. Lucas, L. Vuurpijl, and W. Wiegerinck, eds., pp. 59–68, SNN, Radboud University Nijmegen, 2003.
15. G. Derefeldt, T. Swartling, U. Berggrund, and P. Bodrogi, "Cognitive color," *Color Research & Application* **29**(1), pp. 7–19, 2004.
16. J. Sturges and T. W. A. Whitfield, "Locating basic colours in the munsell space," *Color Research and Application* **20**, pp. 364–376, 1995.
17. T. Myer, "Card sorting and cluster analysis," tech. rep., IBM developerWorks, 2001.
18. S. K. Card, A. Newell, and T. P. Moran, *The Psychology of Human-Computer Interaction*, Mahwah, NJ, USA: Lawrence Erlbaum Associates, Inc., 1983.
19. P. Wilken and W. J. Ma, "A detection theory account of visual short-term memory for color," *Journal of Vision* **4**(8), p. 150a, 2004.
20. R. A. Rensink, "Grouping in visual short-term memory [abstract]," *Journal of Vision* **1**(3), p. 126a, 2001.
21. A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: a review," *ACM Computing Surveys* **31**(3), pp. 264–323, 1999.
22. P. Berkhin, "Survey of clustering data mining techniques," tech. rep., Accrue Software, Inc., San Jose, CA, 2002.
23. E. L. van den Broek and E. M. van Rikxoort, "Evaluation of color representation for texture analysis," in *Proceedings of the Sixteenth Belgium-Netherlands Artificial Intelligence Conference*, R. Verbrugge, N. Taatgen, and L. R. B. Schomaker, eds., pp. 35–42, (Groningen - The Netherlands), 2004.
24. E. L. van den Broek and E. M. van Rikxoort, "Colorful texture analysis," *Pattern Recognition Letters* , [submitted].
25. J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih, "Image indexing using color correlograms," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, G. Medioni, R. Nevatia, D. Huttenlocher, and J. Ponce, eds., pp. 762–768, 1997.
26. R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification," *Transactions on Systems, Man and Cybernetics* **3**(6), pp. 610–621, 1973.
27. M. Sharma and S. Singh, "Evaluation of texture methods for image analysis," in *Proceedings of the 7th Australian and New Zealand Intelligent Information Systems Conference*, R. Linggard, ed., pp. 117–121, ARCME, (Perth, Western Australia), 2001.
28. K. Valkealahti and E. Oja, "Reduced multidimensional histograms in color texture description," in *Proceedings of the 14th ICPR*, **2**, pp. 1057–1061, (Brisbane, Australia), 1998.
29. B. Göransson, *Usability design: A framework for designing usable interactive systems in practice.* PhD thesis, Uppsala University: Department of Information Technology, June 2001. ISSN 1404-3203.
30. M. D. Byrne, *Cogntive architecture*, ch. 5, pp. 97–117. Mahwah, NJ: Lawrence Erlbaum, 2002.