M-HinTS: Mimicking Humans in Texture Sorting

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

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

[b]Faculty of Behavioral Sciences, University Twente
P.O. box 217, 7500 AE Enschede, The Netherlands
e.l.vandenbroek@utwente.nl http://users.gw.utwente.nl/BroekEL/

[c]Image Sciences Institute (ISI), University Medical Center (UMC) Utrecht
Heidelberglaan 100, 3584 CX Utrecht, The Netherlands
eva@isi.uu.nl http://www.isi.uu.nl/

[d]Institute for Computing and Information Science (ICIS), Radboud University Nijmegen
P.O. Box 9010, 6500 GL Nijmegen, The Netherlands
T.Kok@student.ru.nl http://eidetic.ai.ru.nl/thijs/
T.Schouten@cs.ru.nl http://www.cs.ru.nl/~ths/

ABSTRACT

Various texture analysis algorithms have been developed the last decades. However, no computational model has arisen that mimics human texture perception adequately. In 2000, Payne, Hepplewhite, and Stoneham and in 2005, Van Rikxoort, Van den Broek, and Schouten achieved mappings between humans and artificial classifiers of respectively around 29% and 50%. In the current research, the work of Van Rikxoort et al. was replicated, using the newly developed, online card sorting experimentation platform M-HinTS: http://eidetic.ai.ru.nl/M-HinTS/. In two separate experiments, color and gray scale versions of 180 textures, drawn from the OuTex and VisTex texture databases were clustered by 34 subjects. The mutual agreement among these subjects was 51% and 52% for, respectively, the experiments with color and gray scale textures. The average agreement between the k-means algorithm and the participants was 36%, where k-means approximated some participants up to 60%. Since last year's results were not replicated, an additional data analysis was developed, which uses the semantic labels available in the database. This analysis shows that semantics play an important role in human texture clustering and once more illustrate the complexity of texture recognition. The current findings, the introduction of M-HinTS, and the set of analyzes discussed, are the start of a next phase in unraveling human texture recognition.

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

1. INTRODUCTION

In the fields of image processing and computer vision, a range of algorithms exist that describe image properties denoted as texture features. However, so far no adequate, computational models have been developed that mimic human texture classification. Even within the framework of a limited set of textures, huge problems arise when mapping human perception to texture analysis techniques.[1,2]

In 2000, at Human Vision and Electronic Imaging V, Payne, Hepplewhite, and Stoneham[1] applied, as they called it themselves, "perceptually-based metrics to textural image retrieval methods". Payne et al.[1] selected a collection of ten artificial texture analysis methods. Solely, none of these methods agreed with human classifications on more than 29% of the retrieved images. Due to the complementary characteristics of the texture analysis methods, combinations of three methods achieved an agreement of up to 52% with their human counter parts.

The research of Payne et al.[1] reflected the state-of-art in mimicking human texture analysis at that time. Moreover, it illustrated that the focus on texture analysis lay on intensity differences, which was common up

Human Vision and Electronic Imaging XI, edited by Bernice E. Rogowitz, Thrasyvoulos N. Pappas, Scott J. Daly,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 6057, 60570X, © 2006 SPIE-IS&T · 0277-786X/06/$15

till then. Only in the latter five years, a shift in research on texture became apparent: more research was conducted toward color induced texture. Nevertheless, intensity based texture remains of importance since many applications that utilize texture analysis rely on intensity differences.

Five years after Payne et al.[1], Van Rikxoort, Van den Broek, and Schouten presented their paper "Mimicking human texture classification"[2] at Human Vision and Electronic Imaging X. Similarly to the work of Payne et al.[1], they compared human and artificial texture classification. Although the percentage of correctly mapped images between human and artificial classification was higher (around 50%) than the percentage achieved by Payne et al.[1], it was still low. Moreover, Van Rikxoort et al.[2] found that no consistent classification of texture on their dataset was present between humans.

An inspection of the data of van Rikxoort et al.[2] revealed that images in the clusters that were mimicked well by the automatic classifiers, showed little variation in color/gray-scale and texture. So, image clusters of which images have the same texture properties like randomness, directionality, and coarseness were classified good.

Although the research of Payne et al.[1] and Van Rikxoort et al.[2] had a similar approach, the study of Van Rikxoort et al.[2] is distinct from the work of Payne et al.[1] and from other previous research. Van Rikxoort et al.[2] were one of the first who directly compared color induced texture classification with intensity-based texture classification. This was done on the same set of images by the same group of humans and with the same texture analysis methods. Hence, the influence of color on texture classification could be identified.

In the current line of research, the research of Van Rikxoort et al.[2] is replicated: The same experiments are conducted, using the same data, as is explained in Section 3.1. The subjects that participated in the online experiments are introduced in Section 3.2. In the research of Van Rikxoort et al.,[2] the subjects had to meet the experimentator to participate. For the current research it was our aim to i) reach more participants in ii) a more flexible manner, and iii) to enable the parallel participation in experiments by multiple subjects. Therefore, the online M-HinTS card sorting system was developed, as will be introduced in Section 3.3. All experiments were conducted within this newly developed online system. A thorough evaluation was conducted of the system and the experiments conducted with it, as is presented in Section 4. For the latter, the emphasis lay on the difference between the gray scale and color texture sorting experiment, with respect to the execution time of them.

Two distinct analysis are conducted on the gathered data. In Section 5 the consensus between humans and humans and automatic clustering is analyzed. This analysis is adapted from van Rikxoort et al..[2] Section 6 presents a new analysis to inspect the clusters made by the participants and to make a first step in unraveling the relation between semantics and low level visual features. First, we will describe the process of data analysis in Section 6.1. Next, we will describe the findings based on this analysis in Section 6.2. We end this paper with a discussion in Section 7 in which we will address some of the problems encountered and describe follow-up research.

2. AUTOMATIC TEXTURE CLUSTERING

Automatic texture clustering is done in three steps, for both 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.

In this research, three different feature vectors were used for the automatic clustering. In previous work of Van den Broek and Van Rikxoort[3], the optimal set of features for both color and gray scale texture classification were determined. The optimal features for colorful texture classification are four texture features (i.e., entropy, inverse difference moment, cluster prominence, and Haralick's correlation) from the color correlogram[4], based on the 11 color categories[5–8] and the 11 color histogram. For gray scale texture analysis, the optimal features are the four texture features from the co-occurrence matrix[9–11] based on the HSV color space using 32 bins, and a histogram from the HSV color space quantized in 27 bins.

In this set of experiments, 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-scale**					
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

3. M-HINTS: ONLINE HUMAN TEXTURE CLUSTERING

3.1. Experimental setup

The experimental setting used for the current research was taken from Van Rikxoort et al..[2] It was applied on both clustering experiments, using gray-scale and color textures. The data consisted of a collection of 180 colorful texture images were drawn from the OuTex[12,13] and VisTex[14,15] databases. The collection is presented and can be downloaded through: `http://eidetic.ai.ru.nl/M-HinTS/data/`. Two criteria were used when selecting the images for this collection:

1. There had to be images from at least 15 different categories and
2. 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.

In the two experiments, the original color images and their gray-scale counterparts were used. The later set was obtained using the following gray-scale conversion[16]: $I = \frac{R+G+B}{3}$, where I denotes Intensity or gray-value. Now, two identical sets of images were present, except for presence versus absence of color information.

The original problem was how to determine similarities of images within a considerable amount of images. For the latter purpose human ability to cluster data, by way of sorting it manually was exploited. Given a set of texture image, card sorting seemed a useful paradigm for this purpose.

The task of card sorting first appeared as the Wisconsin Card Sorting Test (WCST), which was originally developed to measure conceptual level of the normal adult population[17]. Soon, the test became very popular in clinical neuropsychology, and became part of the basic test battery of a neuropsychologist[17,18].

In the last two decades, card sorting has been applied in broader range of contexts on various types of data[19,20]. Since clustering of images can be seen as sorting the images in a number of categories or stacks, it can be treated as a card sorting task. In the current research, the images contain photographs of natural textures that from now on will be denoted briefly as textures. Hence, the clustering of these textures can be treated as a card sorting. Participant are asked to sort the textures and put them on separate stacks.

The consequence of using stacks of textures is that only the upper texture is completely visible. In addition, some textures are partly visible. During the experiments, the participants had to rely more and more on their memory since both the number of stacks increases and the amount of images on each stack increases. With the increase of the images on a stack, the meaning of the stack also changes; i.e., the representation of the stack in mind. The constant updating of this representation of the stacks implies a constant workload of the participants' visual Short Term Memory (vSTM)[21], which has to be taken into account.

On average human' vSTM can contain four[22]–fourteen[23] items. We, therefore, decided that the number of clusters made by the participants needed to be within this range. Another constraint was that to be able to compare the clusters of textures made by the participants, they all had to define the same number of clusters. The task and data of the current research was similar to the task and data used by Van Rikxoort et al.,[2] who determined the number of clusters to be six; hence, the same number of clusters was taken for the current research.

Table 2: Collected information concerning the computer facilities the participants conducted the experiments on. Please note that when the minimal specifications were not satisfied by the participants computer facilities (i.e., Internet Explorer 4.x or equivalent, screen resolution of 1024×768, and 16 bits color depth of the screen) the participant was not able to continue after the registration phase.

Type of information	Setting	Percentage (%)	#users
Operating System	MS Windows XP®	58.82%	20
	MS Windows 2000®	17.65%	6
	MS Windows 98®	8.82%	3
	Linux®	11.76%	4
	Apple - Mac OS X	2.94%	1
Browser	MS Internet Explorer® 6.x	67.64%	23
	Firefox®	17.65%	6
	Netscape 6/7.x®	11.76%	4
	Apple - Mac OS X: Safari®	2.94%	1
Resolution screen	1280×1024	44.12%	15
	1280×800	8.82%	3
	1152×864	14.70%	5
	1024×768	32.35%	11
Color depth screen	32 bits	73.53%	25
	24 bits	5.88%	2
	16 bits	20.59%	7

3.2. Participants

At the start of the experiment, subjects were required to register themselves. The registration procedure included a questionnaire concerning some personal details: name, gender, education, handedness, glasses, experience with color, colorblindness, age, Email, and computer experience. The resulting data provided the characteristics of the group of participants. Moreover, it enabled us to determine to what extend this group is a representative group of people.

34 people participated of which 20 were men and 14 were women; none of them were colorblind. The average age of the participants is 31.59 (ranging from 12 to 58), with a standard deviation of 12.00. 91.18% of the participants were right-handed, which equals the average percentage of all people (on earth). 44.44% of the participants wore glasses or lenses. The education level of the participants is high; 76.47% has a high or university education. The latter is due to the fact that multiple friends and collegues participated in the experiments. Color experience could be related to that, although this kind of knowledge can be gathered from various sources. 50.00% of the subjects indicated that they have either some and 17.47% even stated that they have much experience with color. As was expected, most participants did have a considerable amount of computer experience; 67.65% has more than 10 years of experience and 17.47% has between five and ten years of experience. However, again should be noted that the high percentage will probably be due to the selected group of participants.

In addition to the questionnaire, the registration phase incorporated the collection of some information concerning the computer facilities of the participants (see Table 2), without that the participants were aware of this. This information incorporated: the operating system, the browser, the resolution of the screen, and the color depth of the screen used.

85.29% of the participants did run a MS Windows® operating system. 67.64% combined it with the MS Internet Explorer® 6.x browser, where 29.41% used either Firefox or Netscape as their browser. Since the window in which the experiments were conducted has a fixed size of 980×650 pixels (see Figure 3.3a), only screen resolutions of 1024×768 and higher were allowed. The most used resolution in this experiment was 1280×1024. The color depth was in a large majority (73.53%) of the settings 32 bits.

3.3. The online M-HinTS experimentation platform

M-HinTS was intended to be accessible for a diverse audience; i.e., subjects had to be able to perform the experiment at home, at work, or at any other PC that was at their disposal. The requirements for the experiment were distilled into the following specification:

- The ability for subjects to register themselves before participation. The data is dealt with confidentially; some aspects can be used for evaluating test results.
- Presentation of instructions; additionally, a preview of a similar procedure is shown.
- The ability to present a randomized set of cards (images) that can be dragged to different positions.
- Storing all the positions (i.e., in x–y coordinates) of the final card layout.
- Visualization of the results of participants on behalf of visual inspection by the experimentators. For the latter purpose, the image IDs and their positions at the end of the experiments are used. See Figure 3.3b for a screendump of the inspection facility.

Every computer is fitted with different hardware (e.g., monitor(s) and input devices) and software, which makes it hard to develop generic applications; especially, when the experiment should be displayed and handled in the same way on every computer. Developing platform independent applications has been made possible by using virtual machines and frameworks (such as Sun Java); however, the required software is not always installed.

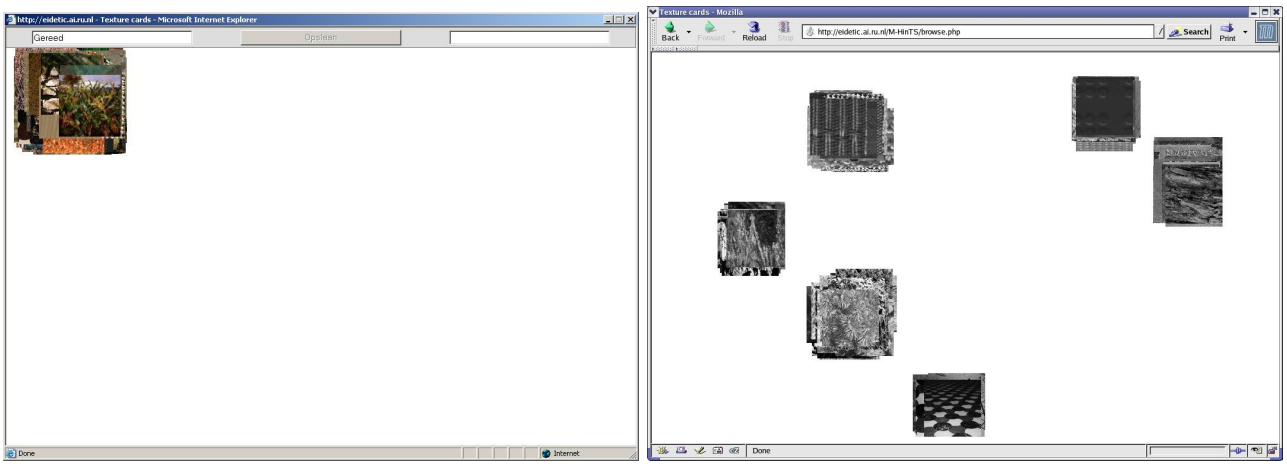
Initially, it was decided to make use of a website, assuming nowadays most computers have a browser installed and are connected to the Internet. Participants would be able to navigate to the M-HinTS website and perform the experiment in their favorite browser. While browsers facilitate ways to use 3rd party utilities such as Java or Flash to enhance functionality, such methods were avoided since these applications are not always installed or compatible with any browser; moreover, different versions of these plug-ins exist, presenting even more problems.

The final implementation of M4ART was build using several techniques to ensure that most computers can be used to access the experiment:

- PHP: The PHP hypertext processor is installed at the webserver and provides efficient methods for dealing with forms (i.e., during the registration process, processing of images, during the card loading procedure) and storing data; i.e., saving the card positions. The interpreter runs at the host computer that facilitates the experiment, which does not require any special functions at the client PC.
- HTML: The HyperText Markup Language is used for the layout of the website. In order to ensure that the website is shown in the same manner in every browser, the M-HinTS system is compatible with W3C HTML 4.01 standard. The visualization of cards is implemented by use of "layers". Each card is represented by a separate layer; every layer is stacked, in the same fashion one would stack real cards. A typical stack would be encoded in HTML as:

  $< \text{body} >$
  $< \text{div id} = \text{`card1}' \text{ style} = \text{`width} : 100\text{px; height} : 100\text{px; top} : 20\text{px; left} : 0\text{px; z} - \text{index} : 1' >$
  $< \text{div id} = \text{`card2}' \text{ style} = \text{`width} : 100\text{px; height} : 100\text{px; top} : 20\text{px; left} : 100\text{px; z} - \text{index} : 2' >$
  $< \text{div id} = \text{`card3}' \text{ style} = \text{`width} : 100\text{px; height} : 100\text{px; top} : 50\text{px; left} : 120\text{px; z} - \text{index} : 3' >$
  $< /\text{body} >$

  Note the specification of dimension, the use of coordinates, using the 'top' and 'left' parameters, and the stack order, using the 'z-index' parameter.
- JavaScript: The JavaScript language is used to facilitate the dragging of the layers (perceived as cards) and a correct stacking visualization; e.g., raising a card when picked up. The scripting language is widely supported by various browsers, although different versions exist. The code has been made suitable for the most common version of JavaScript.



(a) (b)

Figure 1: (a) A screendump of a window of the M-HinTS platform during the start of an experiment. All 180 images randomly placed on one stack in the upper left corner of the window. M-HinTS can be visited through `http://eidetic.ai.ru.nl/M-HinTS/` . (b) A typical end situation of an experiment. All 180 images are assigned to a stack, by a participant. In total, six stacks are present. M-HinTS provides the means to visualize the results of participants on behalf of visual inspection by the experimentators. For the latter purpose, the image IDs and the coordinates of their position at the end of the experiments are used.

4. EXECUTION TIMES EXPERIMENTS: GRAY VERSUS COLOR

At a first glance, no clear differences are present between the experiments with the gray and color texture images with respectively average durations of 655.53 ms and 648.53 ms, which denotes a difference of 7.00 ms; see also Table 3. In addition, no large difference is found between the average duration of the experiments done in gray scale–color and color–gray scale order (678.28 ms versus 622.50 ms; difference: 55.78 ms); again, see also Table 3.

The latter, coarse, first analysis did not reveal any differences between both experiments nor between both orders of execution of the experiments. However, the specification for both experiments per execution order, as is also provided in Table 3, illustrates otherwise. The average values concealed the large differences between each of the four order-experiment combinations.

The differences ($|\Delta|$) between each of the four experiment–order combinations, as denoted in Table 3, indicate the large difference between each of the four combinations. Consequently, the interaction between experiment and order was analyzed using a Repeated Measures ANOVA. The latter analysis of the experiments explained the difference between both experiments ($F(1,32) = 11.00$, $p < .002$). Hence, a strong difference in duration times between the experiments is present, after a correction for the order of execution of the experiments: the gray scale textures were sorted significantly faster than the color textures. In addition, please note the large standard deviations (SD) compared to the average values, for each of the experiments and for both orders of execution, as provided in Table 3. These high SDs indicate strong differences between participants in the time needed to execute the experiments.

5. DATA ANALYSIS 1: SEEKING CONSENSUS

5.1. Method of data analysis

For both experiments, the same data analysis was applied. In this section, the data analysis is described; in the next sections, the results are presented based on this analysis. We will provide a generic definition of our method

Table 3: The average duration (in ms) of both experiments (gray scale (G) and color (C)), specified for both orders (i.e., G–C and C–G). Moreover, the following statistics (all in ms) are provided: i) The absolute differences ($|\Delta|$) between both experiments and between both orders of execution. ii) The averages and iii) standard deviations (SD) for both experiments and both orders of execution. iv) The differences between the averages of both experiments and the averages of both orders.

		experiment		statistics		
		G	C	$\|\Delta\|$	average	SD
order	G–C	748.67	607.89	140.78	678.28	290.96
	C–G	550.53	694.25	143.50	622.50	208.43
statistics	$\|\Delta\|$	197.92	86.36		55.78	
	average	655.53	648.53	7.00		
	SD	310.68	227.60			

of data analysis and, therefore, denote the number of participants as $\#p$ and, subsequently, denote the number of unique pairs of participants $\#(p_i, p_j)$.

For each of the $\#(p_i, p_j)$ ($\#p!/((\#p-2)!\cdot 2!)$) unique pairs of participants (p_i, p_j) a consensus matrix ($M_{(p_i,p_j)}$) of size $c \times c$ was determined, which contains for each pair of clusters (c), 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, the average consensus on the clustering between participants was determined.

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, $C_{(p_i,p_j)} = \sum_{i=1}^{c} max\{row_i\}$. Now, the overall consensus can be determined by: $\sum_{p_i p_j} C_{(p_i p_j)}/\#(p_i, p_j)$.

5.2. Results: Among humans

The average consensus between the participants with respect to colorful texture clustering was 51%, ranging from 31% to 72%. The average consensus between the participants with respect to gray-value texture clustering was 52%, ranging from 33% to 68%.

For both color and intensity induced texture clustering, the results obtained by van Rikxoort et al.[2] could not be reproduced. Last year, the average consensus between the participants for color and intensity induced clustering were 57% and 56%. The average consensus between the automatic clustering and the human clustering reported last year was on average 10% better.

5.3. Results: Automatic versus human texture clustering

For the colorful textures, three configurations (i.e., feature vectors) for k-means clustering were used (see Section 2): (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 5.1. The average consensus between human and automatic clustering using only color, only texture, and their combined information was respectively: 35%, 34%, and 36%.

For the gray-scale textures, three configurations (i.e., feature vectors) for k-means clustering were used (see Section 2): (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 5.1. The average consensus between human and automatic clustering using only gray-scale, only texture, and their combination was respectively: 35%, 35%, and 36%.

5.4. Conclusion

Given the results reported in the latter two sections, it is safe to state that the results as reported by Van Rikxoort et al.[2] were far out of reach. A straightforward conclusion is that the task of clustering textures is hard, probably too hard in such a loose setting. Another method of analysis was chosen that enables a inspection on semantic level as well as on low level visual feature level. This analysis uses the semantic categories, as defined in the databases. The next section describes this analysis.

6. IDENTIFYING CLUSTERING STRATEGIES: SEMANTICS VERSUS LOW LEVEL VISUAL FEATURES

6.1. Data analysis 2: Inspection

This subsection describes a new analysis. In the next subsections, the results of this analysis will be discussed. For both the gray scale and the color experiment, the same analysis was conducted, which consists of the following processing steps:

1. The data collected with the experiments was parsed, as follows:

   ```
   foreach participant
       foreach cluster
           determine the images present in that cluster
   ```

   This resulted in a 2×180 matrix for each participant.

2. Representation of image relations:

   ```
   foreach participant
       foreach image
           determine with which image(s) it was put in a cluster
   ```

   This resulted in a binary 180×180 correlation matrix for each participant.

3. All correlation matrices of the individual participants were summed.
   This resulted in one global 180×180 correlation matrix.

4. To determine the six 'average' human clusters, K-means clustering was used. K-means clustering is an iterative process in which K (6 in this research) clusters are defined. The clustering starts by selecting cluster centers; next, each point is added to the cluster with the smallest distance to the center point. After this first iterations, new cluster centers are defined and the process is repeated. In the current research the following parameters were used:

   - distance measure: squared Euclidean distance
   - random selection of seeds
   - 100 iterations

5. Each texture image was drawn from the OuTex database or from one of 14 categories of the VisTex database. Next, we determined for all images in each of the 6 clusters, from which category they came from. The latter was done as follows:

   ```
   foreach cluster
       foreach category
           determine the amount of images
   ```

   This resulted in a 6×15 cluster-category matrix, in which each of the 180 images was represented.

6. Last, the contribution of each category to each cluster, relative to the size of that cluster was determined. The resulting relative cluster-category matrix is graphically represented in respectively Figure 2a and Figure 2b, for respectively the gray scale and the color experiment.

6.2. Results: Describing and explaining the cluster

The analysis described in the latter subsection enabled us to inspect the clusters with respect to both low level visual features and semantics. For the gray scale textures, Figure 2a illustrates that most clusters are dominated by the texture images from a limited number of categories (< 6). Where some are mostly defined by one (i.e., cluster 1) or two (i.e., cluster 3) categories, the other clusters (i.e., clusters 2, 4, 5, and 6) are more diverse in content. For the color textures, Figure 2b illustrates that most clusters are dominated by the texture images from a limited number of categories (< 4). Where some are mostly defined by one (i.e., cluster 6) or two (i.e., cluster 5) categories, the other clusters (i.e., clusters 1, 2, 3, and 4) are more diverse in content.

The semantics of the labels of the categories of the texture images can be used to verify whether or not the participants sorted the texture images based on their semantics or not. Van Rikxoort et al.[2] suggested that the influence of semantics would be larger with the color textures than with their gray scale counterparts, since color is an important feature in recognizing texture.[2,3] Let us verify this claim by describing the clusters found, using the same 15 categories. After that we will compare the sets of gray scale with the set of color clusters. .

First we will explore the clusters of gray scale textures. Cluster 1 is described by bark, sand, stone, and fabric. On the one hand, one can state that this are all natural materials; on the other hand, their textures are structured and detailed. Cluster 2 is mainly defined by paintings and terrain. For this cluster the only plausible explanation seems their texture that is irregular. Cluster 3 is completely described by OuTex and food textures. For the latter, no straightforward explanation can be given based on semantics; in contrast, the fine, regular dot-like texture structure can explain the combination of both categories. Cluster 4 is mainly defined by OuTex textures; hence, whether this is done based on low level visual features or on the semantic value of the textures can not be verified. Cluster 5 is dominated by the categories leaves and flowers. So, the content of cluster 5 seems to be based on the semantic concept flora. An explanation for cluster 6 is hard to find based on low level features. The content of cluster 6 can be explained by a label such as man-made texture patterns.

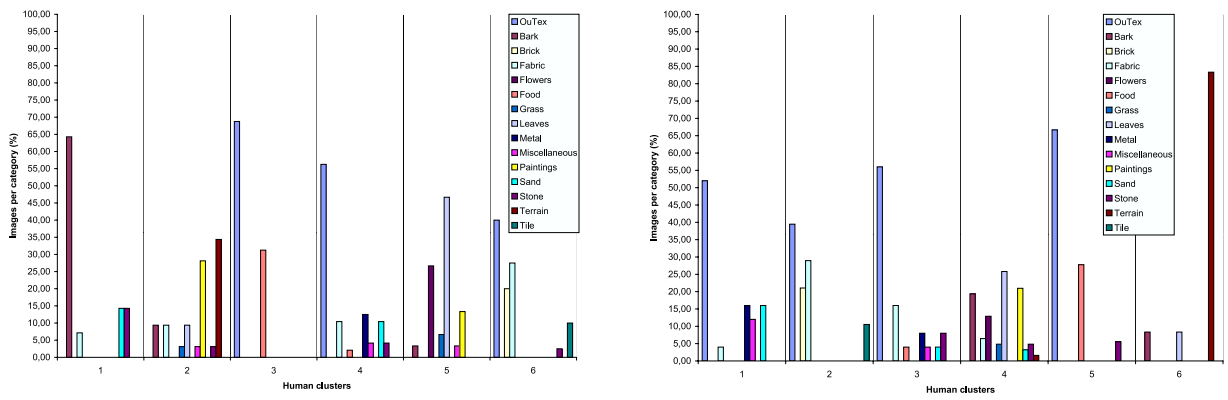
The first color cluster is described by OuTex, metal, miscellaneous, and sand. No strong semantic relation is present between these categories. They have, on the other hand, all a regular, fine texture; hence, it seems most likely that the latter determined this cluster. The description of cluster 2 is almost identical to that of cluster 6 of the gray scale textures; so, the same explanation can be adopted: it are all man-made texture patterns. Cluster 3 is mostly described by OuTex textures; so, no further conclusions can be drawn with respect to semantics versus low level visual features. Cluster 4 is possibly the garbage cluster, as was already mentioned in Van Rikxoort et al.[2]. Textures from eight categories are present in this cluster; the most prominent are: bark, flowers, and paintings. The common ground of the latter three categories is hard to describe from either a semantic or a low level visual point of view. Cluster 5 is dominated by OuTex textures and completed by food and, in a lesser degree, stone. Again neither from semantic nor from low level visual perspective a straightforward explanation can be given. The category terrain dominates the sixth cluster. Whether this is due to semantics or low level visual features can not be determined.

A comparison of the gray scale and color textures and, subsequently, some general pointers towards human clustering strategies is impossible, given the data collected with the two experiments. However, what is clear is that the claim of Van Rikxoort et al.,[2] who suggested that the influence of semantics would be larger with the color textures than with their gray scale counterparts, can not be confirmed. This despite the fact that without any doubt color is of significant influence. Whether color influences the low level visual percepts, the recognition of semantics, or both could not be determined. Considering the descriptions of the clusters provided in the previous two paragraphs, it seems most plausible that color influences both.

6.3. Results: Automatic versus human clustering

In order to be able to compare the automatic clustering to the human clustering based on the semantic categories, cluster-category histograms were calculated for the automatic clusters, see Figure 3. These histograms are constructed by counting the number of occurrences of images from each category in each cluster and dividing it by the size of the cluster, and thus show the contribution of each category to each cluster.

For the gray-scale textures, Figure 3a shows that in the automatic clustering based on global gray-scale and gray-scale texture features, two clusters (i.e., cluster 1 and 2) solely consist of OuTex images. The other



(a) (b)

Figure 2: For both the gray scale (a) and color experiment (b), the specification of the 'average' clusters as made by the participants in terms of the 15 categories the images belong to.

four clusters contain images from diverse categories, 8.75 categories on average. Where in the human gray-scale clustering most clusters were dominated by one or two categories, the automatic clustering contains great variance in three (i.e., clusters 3, 5, and 6) clusters.

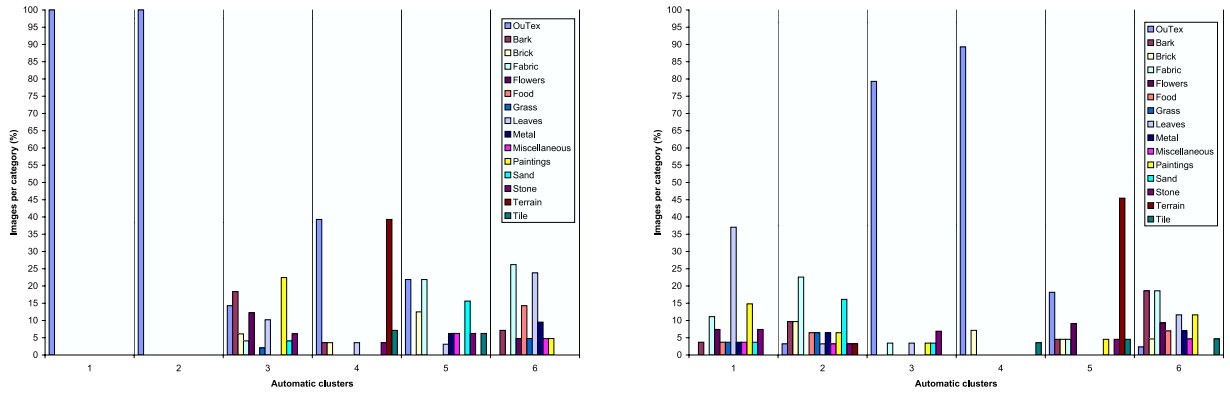
For the color textures, Figure 3b shows that in the automatic clustering based on global color and color induced texture information, the mean number of categories is 8.80. Two clusters (i.e., cluster 3 and 4) are dominated by OuTex images, the other clusters are very diverse. The main difference with the human clustering is the mean number of categories per cluster, for the human clusters the mean number of categories is 5.00 where the mean number of categories per cluster for the automatic clustering is 8.80. In addition, in five out of six human clusters there was a tendency towards one or two categories, in the automatic clustering this is only the case in two clusters.

7. DISCUSSION

In the present research, first 180 texture images were clustered by a k-means clustering algorithm, using three different feature vectors for both color and texture. Next, 34 humans clustered the same sets of images by means of the online card-sorting system M-HinTS. The mean consensus between humans for both color and gray-scale was determined, as well as the mean overlap between humans and automatic clustering. Because the results presented by van Rikxoort et al.[2] could not be reproduced, another data analysis was applied to determine the influence of semantics on human texture clustering.

For both colorful and gray-scale textures, little consensus was present between the participants (51% and 52%). The low consensus between participants confirms the conclusion of last year's research that texture clustering is not a trivial task. The comparison of the automatic clustering to the human clustering were not successful, on average the comparison was successful for only 35% of the images. These results gave reason to look at the human clustering more thoroughly using semantics. The semantics used were the labels of the images in the dataset.

The new semantic-based data analysis indicates that semantics influence human texture clustering. Semantics influence was seen in half of the gray-scale clusters and 33% of the color clusters. In half of the color clusters and one of the gray-scale clusters, the clustering strategy of the participants could not be determined. Semantic analysis applied on the automatic clusters showed that the 15 categories that were available in the data were more divided over the clusters in the automatic clustering than in the human clustering. However, for both



(a) (b)

Figure 3: For both the gray scale (a) and color textures (b), the specification of the automatic clusters, based on both color/gray-scale and texture features, in terms of the 15 categories the images belong to.

colorful and gray-scale texture classification it was shown in previous research[3] that the sets of features used in this research were optimal for texture classification. Moreover, these features were used in a content-based image retrieval system which was judged as accurate by its users[24]. So, a valid conclusion would be that using artificial color and texture analysis is suitable for image classification and retrieval but does not mimic human clustering strategies.

In our research, we aim at mimicking the outcome of human texture classification, we do not claim that our method mimics the process of human texture classification. The results of this paper indicate that only mimicking the outcome might not be possible without knowledge of the process of human texture classification. To be able to mimic this process, fundamental research should be conducted. One such research is described by Rao and Lohse[25] who developed a texture naming system by asking humans to judge the properties of a set of texture images. They showed promising results, however, their experiment was conducted on a small set of gray scale images. For future research, this type of research should be extended to use more images and both colorful and gray-scale textures. Next, the outcome of this research should be converted to features, which can be used for an experiment as described in this paper.

A potential problem of our approach is the fixed number of clusters. One of the problems of the fixed number of clusters is the "garbage group" in which participants put all images they are unable to label. However, when participants are completely free in choosing the number of clusters, there is a change that they will define a lot of very small clusters, which is not informative anymore. In addition, it is hard to compare participants and automatic clustering when the number of clusters varies. Therefore, in future research, the number of clusters might still be fixed, but there should be one special cluster to put "garbage images" in, the size of this cluster should be limited.

The research toward human texture analysis and classification has just started. In this paper, the first aim was to reproduce the results of previous work using more participants. However, the results could not be reproduced, therefore a semantic based analysis was done, which clearly shows semantics is of importance. So, this paper, although explorative, gives substantial information in understanding human texture clustering.

REFERENCES

1. 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.
2. E. M. van Rikxoort, E. L. van den Broek, and Th. E. Schouten, "Mimicking human texture classification," *Proceedings of SPIE (Human Vision and Electronic Imaging X)* **5666**, pp. 215–226, 2005.
3. E. L. van den Broek and E. M. van Rikxoort, "Parallel-sequential texture analysis," *Lecture Notes in Computer Science (Advances in Pattern Recognition)* **3687**, pp. 532–541, 2005.
4. 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.
5. B. Berlin and P. Kay, *Basic color terms: Their universals and evolution*, Berkeley: University of California Press, 1969.
6. R. M. Boynton and C. X. Olson, "Locating basic colors in the OSA space," *Color Research & Application* **12**, pp. 107–123, 1987.
7. G. Derefeldt, T. Swartling, U. Berggrund, and P. Bodrogi, "Cognitive color," *Color Research & Application* **29**(1), pp. 7–19, 2004.
8. J. Sturges and T. W. A. Whitfield, "Locating basic colours in the munsell space," *Color Research and Application* **20**, pp. 364–376, 1995.
9. 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.
10. 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.
11. K. Valkealahti and E. Oja, "Reduced multidimensional histograms in color texture description," in *Proceedings of the 14th International Conference on Pattern Recognition (ICPR)*, **2**, pp. 1057–1061, (Brisbane, Australia), 1998.
12. University of Oulu, "University of oulu texture database." URL: `http://www.outex.oulu.fi/`, [Last accessed on December 30, 2005].
13. T. Ojala, T. Mäenpää, M. Pietikäinen, J. Viertola, J. Kyllönen, and S. Huovinen, "Outex - new framework for empirical evaluation of texture analysis algorithms," in *Proceedings of the 16th International Conference on Pattern Recognition*, **1**, pp. 701–706, (Quebec, Canada), 2002.
14. Massachusetts Institute of Technology, "Vision Texture." URL: `http://vismod.media.mit.edu/vismod/imagery/VisionTexture/vistex.html`, [Last accessed on December 30, 2005].
15. S. Singh and M. Singh, "Texture analysis experiments with MeasTex and VisTex benchmarks," *Lecture Notes in Computer Science (Advances in Pattern Recognition)* **2013**, pp. 417–424, 2001.
16. Th. Gevers and A. W. M. Smeulders, "Color based object recognition," *Pattern Recognition* **32**(3), pp. 453–464, 1999.
17. R. K. Heaton, *A manual for the Wisconsin Card Sorting Test*, Odessa, FL: Psychological Assessment Resources, 1981.
18. E. A. Drewe, "The effect of type and area of brain lesion on Wisconsin Card Sorting Test performance.," *Cortex* **10**, pp. 159–170, 1974.
19. J. Nielsen and D. Sano, "Sunweb: User interface design for sun microsystem's internal web," in *Electronic Proceedings of the second World Wide Web conference '94: Mosaic and the Web*, 1994.
20. T. Myer, "Card sorting and cluster analysis," tech. rep., IBM developerWorks, 2001.
21. S. K. Card, A. Newell, and T. P. Moran, *The Psychology of Human-Computer Interaction*, Mahwah, NJ, USA: Lawrence Erlbaum Associates, Inc., 1983.
22. 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.
23. R. A. Rensink, "Grouping in visual short-term memory [abstract]," *Journal of Vision* **1**(3), p. 126a, 2001.
24. E. L. van den Broek, E. M. van Rikxoort, and Th. E. Schouten, "Human-centered object-based image retrieval," *Lecture Notes in Computer Science (Advances in Pattern Recognition)* **3687**, pp. 492–501, 2005.
25. A. R. Rao and G. L. Lohse, "Towards a texture naming system: Identifying relevant dimensions of texture," *Vision Research* **36**(11), pp. 1649–1669, 1996.