Simulating Fixations when Looking at Visual Arts

Hermann Pflüger, Benjamin Höferlin, Michael Raschke, Thomas Ertl, Institute for Visualization and Interactive Systems (VIS), University of Stuttgart

Simulating Fixations when Looking at Visual Arts

HERMANN PFLÜGER, BENJAMIN HÖFERLIN, MICHAEL RASCHKE, THOMAS ERTL, Institute for Visualization and Interactive Systems (VIS), University of Stuttgart

When people look at pictures, they fixate on specific areas. The sequences of such fixations are so characteristic for certain pictures that metrics can be derived which allow to successfully group similar pieces of visual art. However, determining enough fixation sequences by eye tracking is not practically feasible for large groups of people and pictures. In order to get around this limitation, we present a novel algorithm which simulates eye movements by calculating scan paths for images and time frames in real time. The basis of our algorithm is an attention model which combines and optimizes rectangle features with Adaboost. The model is adapted to the characteristics of the retina and its input is dependent on a few earlier fixations. This method results in significant improvements compared to previous approaches. Our simulation process delivers the same data structures as an eye tracker and can thus be analyzed by standard eye tracking software. A comparison with recorded data from eye tracking experiments shows that our algorithm for simulating fixations has a very good prediction quality for the stimulus areas on which many subjects focus. We also compare the results with those from Itti and Koch [2013], and Niu et al. [2012]. Finally, we demonstrate how the presented algorithm can be used to calculate the similarity of pictures in terms of human perception.

Categories and Subject Descriptors: I.2.0 [**Artificial Intelligence**]: General—*Cognitive Simulation*; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Perceptual Reasoning*; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis—*Object Recognition*; I.5.1 [**Pattern Recognition**]: Models—*Statistical*; I.5.2 [**Pattern Recognition**] Design Methodology; I.5.4 [**Pattern Recognition**]: Applications—*Computer Vision, Signal Processing*; I.6.5 [**Simulation and Modelling**] Model Development

General Terms: Algorithms, Experimentation, Human Factors

Additional Key Words and Phrases: Visual attention, eye movement, perception

ACM Reference Format:

Hermann Pflüger, Benjamin Höferlin, Michael Raschke, and Thomas Ertl. 2014. Simulating Fixations when Looking at Visual Arts. *ACM Trans. Appl. Percept.* V, N, Article (January YY), 20 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

Our aim is the development of techniques which facilitate the search, exploration, and general structuring of large collections of works of visual arts. To this end we defined a feature space with an associated function that maps every image to a unique point in the feature space (cf. "analytic aesthetics" in Nake [1974]). We want to use this construction to put a large number of pictures in relation to

Author's address: Hermann Pflüger, Universitätsstraße 38, 70569 Stuttgart, Germany; email: Hermann.Pflueger@vis.uni-stuttgart.de; Benjamin Höferlin, email: Benjamin.Hoeferlin@vis.uni-stuttgart.de; Michael Raschke, email: Michael.Raschke@vis.uni-stuttgart.de; Thomas Ertl, email: thomas.ertl@vis.uni-stuttgart.de

DOI: http://dx.doi.org/10.1145/0000000.0000000

their visual impact. We started out by defining a sign system which takes advantage of the fact that pictures designed by artists are related to mental states. Pictures trigger certain mental states, and mental states, in turn, may be expressed by suitable pictures. Working on the assumption that pictures that are similar in human terms invoke similar associations, we defined a sign in our sign system as a set of images that are similar to each other and therefore trigger similar associations. Based on this sign system, we further defined the feature space with an associated mapping function. Each sign in our sign system represents one dimension of this space, and the coordinate value of a given image for each dimension is the measure of similarity between the image and the respective sign. As the ability to determine similarity in human terms is crucial for this approach, we had to develop an algorithm that would enable us to do so.

A state-of-the-art method of analyzing images with respect to their similarity is defining characteristic areas, e.g., salient regions, and descriptors for these areas, e.g., local histograms. The regions with their descriptors result in features which are representative of the respective images. If two pictures have matching features, the images are considered to be similar (Aly [2011], Mikolajczyk et al. [2005]). Central to the comparison of images by means of this method is the determination of appropriate areas in the pictures. Here, mathematical methods aimed at identifying points with high salience like edge detectors are typically used, and the surroundings of these points are taken as characteristic areas of the respective image. However, this approach has two disadvantages. First, it takes about one thousand characteristic areas per image to find an image in a large collection of images with a sizeable measure of reliability (cf. Aly [2011]). This results in large memory requirements and high computation times. Second, while the usual methods to identify characteristic regions take account of visual conspicuousness, they do not consider what people actually perceive with their retinas and what is important for human perception. Thus, the methods calculate similarity more theoretically than in human terms. Hence we wanted to find a way to discover areas in pictures that characterize the latter on the basis of a smaller number of those areas and which more closely approximate human perception.

There are many experiments which show that focal attention, visual perception, and fixations are strongly correlated (e.g. as discussed by Hoffman and Subramaniam [1995], Saarinen [1993], Niu et al. [2012]). We also know from our own experiments that one can quite well distinguish images of visual art simply by using the distribution of fixations. Preliminary studies show that 50 to 100 fixations in pictures are enough for that purpose. Therefore we assume that fixations determine those characteristic regions that we are looking for. However, according to our knowledge, there are as yet no algorithms which can simulate realistic scan paths with fixations in sufficient number and which is also fast enough to compute a large number of images within a reasonable time period.

Of course, fixations which are characterized by complex cognitive processes or which are caused individually cannot be simulated. However, based on preliminary experiments, we know that about 70% of the fixations of one participant were also fixated by over 50% of all participants, and hence could be interpreted as not being individually motivated or chosen by complex cognitive processes. With the algorithm presented in this paper we can identify most of these 70% of not individual motivated fixations (Section 4.2). Another algorithm presented here calculates similarity between pictures using the distribution and surroundings of fixations as features. This approach has the advantage of taking not only the visual saliency of regions into account, but also the way a number of observers weight different salient areas of the images with respect to these areas' importance for the perception of a picture. First results from a prototype implementation of an algorithm based on this approach (Section 5) show that the calculated similarity can be interpreted as being human-like. It also shows good results when it is applied to abstract pictures, which are difficult to analyze for this purpose.

In recent years, the increased sophistication and accessibility of eye tracking technologies have generated a great deal of interest in using eye tracking data in a wide variety of disciplines, as for example in cognitive science, psychology, psycholinguistics, and human-computer interaction. There are many software programs available to visualize and analyze such data. This evaluation software can be used together with our simulation algorithm for fixations, because it delivers the same kind of data which an eye tracker does, and although it takes no great effort to acquire eye tracker data for a given image, there are important applications where eye tracking is not possible. Examples include the analysis of a large number of images, automatic optimization of computer-generated images when rendering virtual objects, automatic optimization of computer-generated diagrams, or analysis of eye tracking data during recording. We therefore consider our simulation algorithm for fixations useful far beyond the scenario presented in this paper, particularly since in most applications, it is not individual behavior that is being investigated, but rather average behavior, a fact which is highlighted by our method.

The methods described in this work have been developed, optimized, and tested in this specific context, and the evaluation of the algorithms was done by means of pictures of visual arts. Therefore, the results cannot be generalized for all types of digital images, although the algorithms described in this work, both for the simulation of fixations as well as for the calculation of similarity, are generally applicable. In order to support this assumption, we performed some preliminary testing with various diagrams and tree maps. In these experiments, participants had to evaluate the respective visualizations. Although the algorithms were trained with our corpus of visual arts, and the goal was to get insight into cognitive performance, the simulated fixations reflected the recorded fixations quite well. In particular, the fixations many subjects focused on were among the simulated fixations. Therefore, we assume that our algorithms can be used in other domains, but should be trained with appropriate data and should be evaluated taking the respective objectives into account.

2. MODELS AND SIMULATION WORKFLOW

The interactions between focal attention, visual perception and fixations, and the way they are influenced by visual stimuli and cognitive/affective factors, are extremely complex. So far, we only understand the first steps of seeing in some detail, i.e., the optical system, how the sensor cells in the retina work, and the type of data compression in the retina (see e.g. Rodieck [1998]). Our knowledge of the underlying processes is as yet spotty. Although we know which areas of the cerebral system are involved in the process of visual perception and how they are linked, the principles of how visual stimuli and cognitive/affective factors determine the focus of attention and fixations, respectively, can only be approximated on the basis of experiments. Therefore, our approach is to create a model of the retina based on the current neurophysiological findings while using a learned model by means of eye tracker data to calculate how the information from the retina determines the fixations.

The basis of our simulation algorithm for fixations is an attention model which combines and optimizes rectangle features with Adaboost (see e.g. Viola and Jones [2004]). Depending on a few earlier fixations, the input for this model is adapted to the characteristics of the retina. Considering the dependence of the visual input on the location of the fixations is a new approach which strongly improved our results. We extended the attention model to calculate the probability with which each pixel of a given picture would be fixated on next with respect to the previous fixations, which is also a novel approach. Figure 1 depicts the schematic workflow of this process.

2.1 Visual Attention Model

The guided search model by Wolfe [1994] claims that attention is based on visual stimuli (bottom up) as well as on demands of the observer (top down). Jasso and Triesch [2007] explain bottom-up and top-down mechanisms in more detail. Bottom-up mechanisms can be characterized as automatic, re-

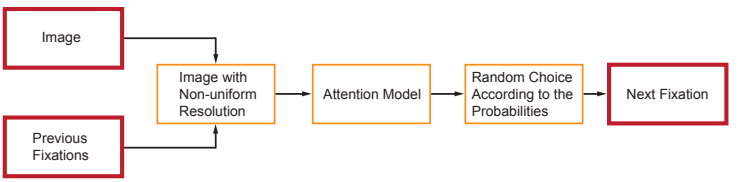

Fig. 1. Schematic workflow for calculating a new fixation

flexive, fast, massively parallel, and requiring only a comparatively simple analysis of the visual scene. They depend neither on the knowledge of the user nor on the search task. Top-down mechanisms are slow and more deliberate, and require complex inferences and the use of memory. According to Wolfe, early vision stages separate the visual stimuli into different feature maps by means of bottom-up activations. Each feature map contains a different feature channel, such as color, orientation, motion, or size; and every feature map is like a heat map that shows how unusual a specific feature is in comparison to its surroundings. Top-down activation, on the other hand, emphasizes the features the observer is interested in (e.g. persons) and combines the feature maps (bottom-up) by means of a weighted sum into a single activation map. The activation map determines which locations receive attention. Other methods use clustering algorithms to find those regions which have a high density of salient points within a large set of salient points (Rajashekar et al. [2008] and Privitera and Stark [2000]). These clusters are presumed to contain the most striking picture elements which receive the most attention. Learned models were also proposed. For instance, Judd et al. [2009] utilize a linear support vector machine to train a model of visual saliency including low-level, mid-level, and high-level features to combine bottom-up signal cues and semantic top-down cues. In this terminology, the low-level features are equivalent to the feature maps above, the high-level features correspond with the points of interest in the aforementioned top-down mechanisms, and the mid-level features are preliminary stages of cognition, e.g., a horizon line detector. Itti [2005] presents an approach to calculate bottom-up saliency of video data. Collecting eye tracking data of subjects and creating saliency maps by means of a computational model that considers low-level features, he finds that motion and temporal features are more important than color, intensity, and orientation. However, the best predictions are achieved by a combination of all these features. Davis et al. [2007] trained a focus-of-attention model to create pathways for pan/tilt/zoom cameras. Their model utilizes a single feature, i.e., translating motion, to capture the volume of activity. Kienzle et al. [2007] trained a feed-forward neural net. In their approach, the video is smoothed spatially and filtered temporally. Training the neural net optimizes the temporal filters along with their weights. Another approach by Nataraju et al. [2009] combines a modified version of Kienzles method with the visual attention model of Itti et al. [1998], the latter of which is based on saliency maps. This approach uses a neural net to train the coefficients of three low-level descriptors (color intensity, orientation, and motion). There are also models that use algorithms for object recognition in order to take top-down effects into account. The method of FG2 [2014] can recognize faces and texts, and can combine such features with bottom-up effects. However, suitable objects must be selected depending upon the particular application. For example, face and text recognition does not make sense applied to abstract paintings or landscapes.

In contrast to the above-mentioned methods, our model is not restricted to a single feature/channel (Kienzle et al. [2007]), to a saliency map from a single feature/channel (Davis et al. [2007]), or to a set of predefined channels (Nataraju et al. [2009]). With this model, the training approach is based on temporal and spatial rectangle features, optimized with eye tracking data, and can thus represent rather arbitrary channels, such as lightness contrast, color contrast, motion, orientation, and sym-

metry. This means that the model does not require manually modeled channels. Here, the bottom-up cues are learned from training data. Furthermore, the contribution of each feature to the final saliency map is determined by the training process. Hence, two important issues with channel-based saliency maps are addressed: the selection of features as well as their weights. This approach also covers top-down mechanisms, as for instance those used by Judd et al. [2009] or FG2 [2014], e.g., face and people detectors. Such high-level features are implicitly learned by the method itself.

The attention model which we introduce in this paper is grounded on a model we had developed for video input (Höferlin et al. [2012]). We approximate the operating principles of the retina (cf. Rodieck [1998]) by means of rectangle features with thresholds. The types of rectangle features were carefully chosen with respect to the causal mechanisms that are known to attract human visual attention. The rectangle features work analogously to the receptive fields of the retina. The types of rectangle features used in our model are shown in Figure 2. By combining and optimizing these features with Adaboost (cf. Viola and Jones [2004]), our model is able to represent complex signal characteristics, such as lightness contrast, color contrast, alterations, orientation, and symmetry. These signal characteristics represent the major cues for attention guidance according to Itti [2005] and Wolfe [1994]. Our approach thus includes the typical categorical channels of bottom-up attention models based on saliency. At the same time, it also solves the problem of selecting the individual weights of each channel by learning their contributions with eye tracker data, whereas other approaches often require the manual assignment of those weights. Furthermore, our approach is able to learn particular channels that have not been defined beforehand. While manually defined saliency operators need an exact definition of such channels, our method only requires a set of features capable of covering these bottom-up issues. In this way, additional channels that are not explicitly mentioned are learned from the data. We do not distinguish between the original mechanisms of the fixation process (bottom-up and top-down). Instead, we are only concerned with the signal characteristics of the image data. At the same time, though, fixations stemming from top-down guidance also affect our attention model. Test subjects focus on high-level objects (e.g. faces, eyes) and mid-level objects (e.g. horizon line, simple geometric objects) more often than on salient regions that are not connected to objects involving cognition (see e.g. Niu et al. [2012]). For example, subjects often fixate on special regions in a face (e.g., eyes, mouth, and ears) displayed in an image. Thus, based on the respective eye tracking data, Adaboost chooses such features which are typical of these regions, assigning them great weights. As a consequence our attention model is sensitive to patterns correlated to faces (cf. Viola and Jones [2004]). In this respect, the attention model we trained is to some extent also an object detector, responding to signal patterns of typical top-down mechanisms.

Compared with current methods, our algorithm scored the best results by far. However, the input of this model is video data, in which context temporal features have proven to be most efficient. Experiments in Höferlin et al. [2012] with videos indicate clearly that the performance of the learned visual attention model depends on the set of features Adaboost used for selection. For instance, a model that includes only simple edge/contrast detectors is actually useless for the prediction of fixations. On the other hand, temporal features describing the change of the lightness channel prove most efficient. However, a combination of all features, including color information, shows overall best performance. These observations are consistent with the results of Itti [2005]. However, the input in our case consists of time-invariant images, so that these most efficient features could not be used directly. The following considerations suggest that these features, and hence the given model, can be adapted to our case: The pictures which were given to participants were pixel images with a uniform resolution. The human eye, however, only gathers detailed visual information in a small central region of the visual field, the fovea. Retinal spatial resolution decreases as the distance to the fovea increases (cf. Figure 3). Therefore an observer of a static image gets a different visual input with every new fixation. This

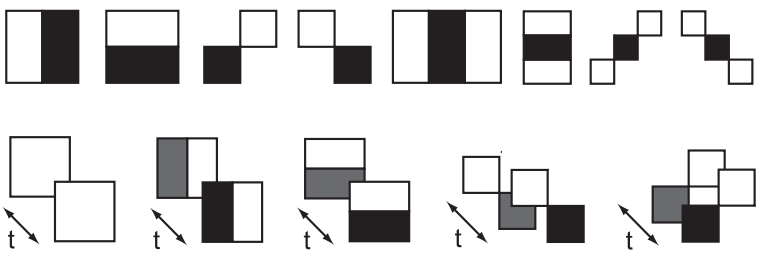

Fig. 2. Basic types of rectangle features used to train the visual attention model. Top row: spatial edge detectors (1-4) and spatial ridge detectors (5-8). Bottom row: Temporal difference operator and spatial-temporal edge detectors.

Fig. 3. Left image: Original Image. Right image: Visual input with most recent fixation (marked with a black and white square) in the face of the woman.

fact is important, because the resolution with which alterations in the periphery are detected is much higher than the resolution for contrast effects. Of course an observer does not perceive motion if visual input changes due to changing fixations. But features such as edges or color also cause a change in visual input if a new fixation comes closer or recedes. This suggests that the detection of a change in visual input in human perception is not exclusively used for the recognition of movements. Our results were much more conclusive when we took this effect into account: With only spatial rectangle features applied to images with a uniform resolution as input, only 25% of recorded fixations lie in regions classified as positive (on average 5% of the whole picture); 28% were achieved when the resolution of the input images was adapted to the fixations and the qualities of the retina; but 62% were achieved when temporal rectangle features were used in addition. Therefore, we used images with non-uniform resolutions and time-dependent rectangle features for the selection and optimization of the features with Adaboost. As a result, the resolution of these images depends on the latest fixations and the neurophysiology of the retina. We calculated location-dependent resolution with respect to the results of Rodieck [1998]. In this manner the visual input is time-dependent (with respect to the fixations of the subjects), like the input given by a video, even though a time-invariant image is viewed.

Rajashekar et al. [2008] describe a method that also considers the non-uniform resolution of the retina. They calculate the dependence of the relevance of features of the recent fixation with statistical methods. Their investigations prove that the relevance of local features to the choice of fixations depends strongly on the distance from the previous fixation. The disadvantage of this method is that it is difficult and costly to automatically optimize their features (e.g. the kind of features and their spatial extent) via Adaboost. Moreover, we think that such an effect depends crucially on known neurobiological properties of the retina and can therefore be simulated by imitation as performed by our model, which also suggests a better approximation.

2.2 Extended attention model

There are a number of visual attention models predicting fixations. Most of these models are based on bottom-up mechanisms and use feature/saliency maps. They combine different feature maps and use competing algorithms among points on these maps to pinpoint a single winning location in the picture next in line for investigation (Itti and Dhavale [2003] describe such an algorithm). Alternatively they use clustering methods to find those regions which have a high density of salient points (Rajashekar et al. [2008] and Privitera and Stark [2000]). In this case, the cluster with the highest density is the winning location. To get fixations, one may either lower the conspicuousness of the winning location with a time function so that after a certain time another region dominates (Itti and Dhavale [2003]) and Rajashekar et al. [2008]) or an order of conspicuousness is calculated for all salient regions (Privitera

and Stark [2000]). In the first case the current winning location yields the current fixation; in the second case, the order of conspicuousness directly provides the sequence of fixations. However, these procedures are only a very simple approximation of the complex perceptual process involved here. For example, only the most salient regions are taken into account for fixation simulations. This results in a small set of fixations for a given picture, i.e., the most probable ones. The example given on Itti and Koch's homepage (Itti and Koch [2013]) shows 5 to 20 fixations per picture. One application of these algorithms is to simulate fixations in videos (Itti and Dhavale [2003]), because here, visual input keeps changing, and thus only the most probable fixations per frame are necessary for a simulation. However, the simulation algorithm we envision will use about 100 fixations per picture: fixations that not only reflect the most noticeable points but also points which define the structure of a picture.

We therefore extend our attention model, so as to be able to calculate fixation in the manner outlined in Figure 1. Given the fixations of the last few time steps, our goal is to establish the probability of being fixated on next for each point in the image. Since we do not know the process of human perception for selecting fixations procedurally, we used a learned model trained with eye tracker data. First we built all 512 combinations of the attention model's 9 classifier cascades. Using a new set of fixations, we then generated time-dependent visual input in the way described above. We classified the regions for potential fixations for each cascade combination. Thus we had 512 disjoint regions for every "frame". (Please note that there is a region in which none of the cascades classifies as positive, and so every pixel of the image lies strictly in one of the 512 regions.) Next, we added up the number of pixels $N_{p,i}$ that were classified as positive, and the number of fixations $N_{f,i}$ for every cascade combination i and the whole period of observation. We took the quotient d_i of these numbers as the fixation density $d_i = N_{p,i} \ / \ N_{f,i}$. We used the extended attention model to generate fixation points. In every time step, based on the fixations of the last few time steps, each cascade combination classifies the pixels of the image. Thus, we get a region classified as positive for each cascade combination. In this manner, every pixel of each image is strictly matched to one of the 512 cascade combinations. For each point in the "frame," the matched cascade combination defines the density of fixation. The next fixation point is calculated randomly, with a random distribution corresponding to the fixation density d_i.

3. TRAINING PROCEDURE

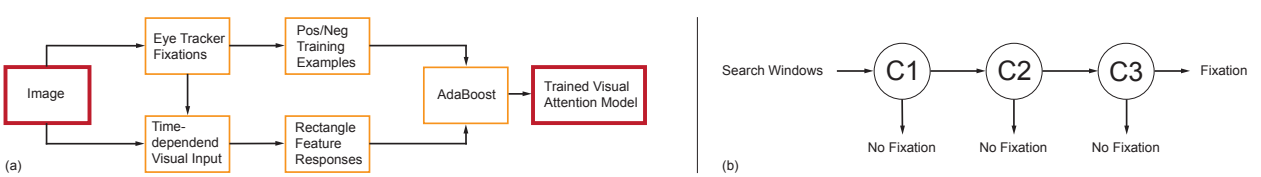

Fig. 4. Schematic workflow of the training process (a). Cascade of three classifiers (b).

We took recorded fixations as positive training examples. Our negative training examples were based on (equally distributed) random points of the visual input, but not within a spatio-temporal suppression radius around fixations. All points with a weighted Euclidian distance $d = \sqrt{dx^2 + dy^2 + a\ dt^2}$ of less than d_{min} from a fixation were ignored. Here dx, dy, dt are the distances of the space and time dimensions (dx, dy measured in pixels, dt measured in time frames). The parameter a used to compensate for the different measures was set at 10, based on previous experiments. The value of the suppression radius d_{min} has a maximum detection rate at 180. Based on the positive and negative training examples and the time-dependent input signals, we created a discriminative visual attention

model consisting of a cascade of classifiers, as depicted in Figure 4. Each classifier consists of a set of rectangle features (cf. Figure 2) selected by Adaboost (cf. Viola and Jones [2004]).

The decision function H of a single classifier (cf. Figure 4b), i.e., the determination of whether a given point is a potential fixation point or not, is calculated by using a weighted linear combination of N (thresholded) rectangle features with their response r_n at the given point.

$$H = \begin{cases} true & \text{if } \sum_{n=1}^{N} \alpha_n \, max(0,\ sign(c_n \ (r_n - t_n - b))) \ \geq \ 0.5 \\ false & \text{if } \sum_{n=1}^{N} \alpha_n \, max(0,\ sign(c_n \ (r_n - t_n - b))) \ < \ 0.5 \end{cases}$$

$$c_n = \begin{cases} 1 & \text{if the response of feature } n \text{ has to exceed the threshold } t_n \\ -1 & \text{else (if the response of feature } n \text{ has to fall below the threshold } t_n) \end{cases}$$

The training of a single classifier includes the selection of the appropriate rectangle features, their thresholds t_n, and their weights α_n. These features were selected by Adaboost from a set of potential weak classifiers, i.e., rectangle features. Finally, we adapted the thresholds of the rectangle features with a uniform bias b. This step directly affected the area which the learned visual attention model marked as a potential fixation area. Lowering the thresholds brought about a generalization of the model and more potential fixations, whereas raising the threshold reduced the number of potential fixations. In order to determine the optimal threshold adaption, we applied a gradient search on the NSS target function (for NSS, normalized scan path saliency, see Peters and Itti [2007]). Experiments suggest that this function has only a global maximum, which can be calculated with a gradient ascent method. There is a drawback using Adaboost for optimizing the classifiers in the classifier cascades. Adaboost can only create and optimize linear combinations of rectangle features (weak classifiers). There are, however, fixations at points where only large rectangle features have a strong response, and there are fixations at points where only small rectangle features have strong responses. If discrimination of any one of these groups is intended, both large and small rectangle features in the classifier have to be used. But such a classifier cannot discriminate points with medium responses for both kinds of rectangle features. The situation is similar with rectangle features for different color channels. In order to take care of this problem, we trained not only one but nine cascades. For each dimension of the L*a*b* color space we trained three cascades that differed in both size and position of rectangle features. Whenever one of the cascades classified a point as positive, that point was then identified as a possible fixation point. One benefit of this approach is that it is not necessary to use differently scaled images to apply the classifier to a given image. A drawback, however, is the way we provide Adaboost with negative training examples. Even if we suppress negative training examples near fixations as described above, this does not guarantee that a selected negative training example may not be a potential fixation, since not all potential fixations can be excluded, precisely because they are not known. However, we do have a high probability that our negative examples are not fixation points. It is thus possible that some of our negative examples had similar signal characteristics as our positive examples. It follows that the linear separability of the training set is not optimal. In Adaboost, classification goals used for boosting and cascade construction were usually predefined ones. Viola and Jones [2004], for instance, take a detection rate of 99% and a false positive rate of 30%. The poor linear separability of our training set renders these classification goals ineligible. Instead, we used a predefined number of classifiers with a predefined number of features each, like Zhao and Koch [2011].

To obtain training data, test persons were shown pictures of visual arts and given the task of memorizing these pictures so that they would be able to recognize them in a subsequent exercise. For this task we assumed that they would pay specific attention to the salient regions that are characteristic regions in the pictures. For detecting the fixations we used a Tobii T60 XL eye tracker. The eye tracker precision was 20 pixels, which is lower than the natural mean deviation of the fixations from the trig-

Fig. 5. Pictures used to receive training data.

Fig. 6. Pictures used to receive data for evaluation.

gering point (Rodieck [1998]). The calibration was tested at the beginning and at the end of every test run to be sure that the precision was constant over the entire test. The first set of positive training data were data recorded from 13 test subjects looking at 11 pictures each (Figure 5). They had 15 seconds to observe each picture, which yielded a total of about 10,000 fixations. Half of the results were used to train the attention model; the other half was first used to evaluate the attention model and then to train the extended attention model (see Section 2.2). Another data set derived from a test with a different group of 14 participants and 10 new pictures (Figure 6) provided new data with about 10,000 fixations. This data was taken to examine the correlation between the fixations of the test persons themselves as well as the correlation between subjects fixations and simulated fixations. Furthermore, the generalization of the attention model was tested based on this new test data.

The pictures presented to the subjects were scaled to a standard diagonal of 0.35 m. The distance from which the participants observed the pictures was 0.6 m. This ratio of size and distance is roughly equivalent to that typically encountered in a museum or while looking at pictures in an art catalogue. The resolution of the screen used was 100 ppi. This conforms with the highest resolution of the human eye (at the fovea) at the viewing distance at hand. The fixations of humans are actually not set at exactly the particular salience that caused the fixation. According to our findings and as confirmed by Rodieck [1998], the fixations seemed to be normally distributed. Both the center and the deviation depend on the distance between the previous fixation and the target fixation (see Rodieck [1998] for details). In order to simplify our calculations, we assume that fixations motivated by salient points are distributed normally around these salient points. The deviation value we use is 32 pixels, which is motivated by several factors: The results in Rodieck [1998]; the fact that a circle on the screen with a diameter of 64 pixels viewed from a distance of 0.6 m corresponds with the area observed by the fovea; and the eye tracker precision of 20 pixels. We assume that a fixation is correctly classified (in the extended attention model) if the results of the following calculations substantiate as much. First, we distribute the probability for the position of the salience that caused the fixation by means of a Gaussian filter in a 32-pixel radius around a given fixation. Then we calculate the value of w with the following equation.

$w = \sum_A (\delta(x,y)\ g(x,y)$ where

(x,y) is a position in the image and

A is a 32-pixel radius around the given fixation.

$$\delta(x,y) \quad = \quad \begin{cases} 1 & \text{if point } (x,y) \text{ is classified as positiv} \\ 0 & \text{if point } (x,y) \text{ is classified as negative} \end{cases}$$

$g(x,y)$ is the impulse response of the Gaussian filter for position (x,y).

If $w > w_t$, we assume that the fixation was classified correctly. We set w_t at a value which ensures that if we proceed as described and if fixations were selected randomly with the same number of positive classified pixels as selected by the attention model, the probability of a random hit was less than 0.00135. The impact of random hits is therefore negligible.

Generating the classifiers of the cascades with Adaboost requires generating a set of rectangle features as a first step. As mentioned above, we used three such sets for every dimension of the L*a*b* color space:

Set 1: All the basic features depicted in Figure 2, modified in size and position in gradations of 4 pixels within a window of 24 x 24 pixels around the pixel that is examined.

Set 2: All the basic features depicted in Figure 2, modified in size and position in gradations of 8 pixels within a window of 48 x 48 pixels around the pixel that is examined.

Set 3: All the basic features depicted in Figure 2, modified in size and position in gradations of 16 pixels within a window of 96 x 96 pixels around the pixel that is examined.

In total, we had 20,832 rectangle features. Out of these, Adaboost chose 135 rectangle features for the 9 classifier combinations. Each combination had 3 classifiers with 5 features per classifier. Because of the problems with the linear separability of the training data mentioned above, more classifiers per combination or more features per classifier did not improve classification.

For our attention model, a constant bias was calculated for each one of the particular cascades (Section 2.1). However, there are pictures with low contrast and pictures with high contrast. The size of the regions that are classified as positive is much smaller in the pictures with low contrast. Hence these regions were underrepresented in the calculation of the fixation density. In order to compensate for this effect and to increase the model's generalization potential, we multiplied the bias of the cascades with a picture-dependent factor (both during the training of the extended attention model and during the simulation process) so that all regions classified as positive covered about 5% of any given image. The value of this factor for all pictures was in a range within which the value of the NSS was close to the maximum. The value of how much of the image should be classified as positive is not decisive. We also used 3% and 9% without significant differences at the simulated fixations. However, it is advisable to use the same value for calculating fixation density as for calculating fixations later.

Surprisingly, there was no discernible dependence between the duration of the fixations and the pictures observed in our tests, which indicates that fixation duration is not directly dependent on visual input signals. Our tests show differences in fixation duration among the participants, but not among the pictures that are observed. The duration of the fixations seems to be Poisson-distributed. The parameter λ depends on the subject and was in a range from 4 to 8. Therefore we implemented a Poisson-distributed duration with the parameter $\lambda = 6$ in our simulation. We also conducted some testing with various diagrams and tree maps. In these experiments, participants had to evaluate the respective graphics. Here, too, we observed the same behavior as above, although in this instance cognition played a part, and specific pixels had to be examined more closely. However, we found that fixation duration with subjects who had to look at videos featuring uneventful sequences of some length was generally prolonged and special image points were fixated particularly long, as if the participants were taking time out. There are also studies providing more detailed examinations of the fixation duration. Holmqvist et al. [2010] describe studies which found associations between visual input and fixation duration especially during reading tasks. However, the evaluation of the duration of fixations and the comparison of the respective results is difficult. Subjects' eyes are never completely fixated on a point; they always move slightly. The definition of what constitutes a fixation depends on the evaluation software of the eye tracker. For example, in our studies, the eyes in portraits were fixated on not particularly long, both in simulation and in the recorded data. But there were always several successive fixations that were close together. If these fixations were considered as a single fixation, this would suggest a relationship between fixation duration and image content.

The distribution of the distance between two fixations depends on the observed picture. We expected that this fact was also learned by means of the training of the attention model. However, the small distances the test subjects tend to go for are not favored to the same degree in the simulation. This does

not significantly affect the distribution of the fixations over a longer time period (about 15 seconds) and is not overly important in terms of our intended applications, but we improved the simulation in this respect by increasing the fixation density in the region near the most recent fixation. The amplification factor was determined experimentally.

4. RESULTS

Real life experience shows that people have a comparable notion of what constitutes similarity of pictures. This observation gives rise to the hypothesis that there are commonalities in perception caused by visual features, and that these features strongly determine the perceived similarity between pictures. It is these features that we set out to detect by calculating fixations during perception. As a matter of fact, our eye tracker data shows that about 70% of the fixations of a given test subject correspond with the fixations of more than 50% of the other subjects, and hence may be interpreted as being not individually motivated. We assume that these fixations determine characteristic regions for perceiving similarity. The following results show that our simulation detects most of these common fixations. In addition, they show that our simulation behaves similarly to human perception in general, so that it can be used to simulate, for instance, the eye movement of avatars, as described by Itti and Dhavale [2003].

4.1 Evaluation of the visual Attention Model

Our visual attention model works with rectangle features selected and weighted by Adaboost (cf. Section 2.1 and 3). There are three issues which affect the procedure. The first issue is the choice of the training data. For positive training data, we used eye tracker data recorded from 13 test participants looking at 11 pictures (data set I; Figure 5). This may seem to be too small a number of pictures with which to train an algorithm that claims to work with any picture of visual arts. In fact, however, as Figure 7a indicates, this number appears to be sufficient. In this instance, the classifiers were trained with 5, 7, 9, and 11 pictures respectively, each time evaluated on the basis of the same data, namely, the 50% of data set I not used for training. The resolution shows a saturation effect. Moreover, another evaluation made with data set II (Figure 6) with different pictures and participants shows that the algorithm's generalization properties are very strong, because the recall rates in this case were even better. The negative training examples are generated with an (equally distributed) random positioning of the visual input. But there is no guarantee that there are no potential fixation points among them, so a spatio-temporal suppression radius is used (cf. Section 3). The second issue is the choice of the set of rectangle features in order to generate the classifiers of the cascades with Adaboost. Based on previous experiments, we used the sets described in Section 2.1. Figure 7b shows the types of rectangle features Adaboost used for creating the cascades. The third issue is both the design and combination of the cascades, for which there are innumerable possibilities. In our evaluation, the extended attention model detects 62% of all fixations while classifying only 5% of each picture as positive. This is a good result, especially since all the test persons and pictures of the test data set were different from those of the training. We therefore refrained from further studies involving different designs of cascades.

Niu et al. [2012] examined the influence of cognitive/affective effects on observers' attention. They compared the number of fixations in regions with both visual and emotional salience, showing as a result that regions with high emotional salience predict fixations much better than regions with high visual salience. They counted the number of fixations within a time range of 2 seconds; for regions with high emotional salience they received about 2 fixations. With our extended attention model we were able to improve prediction accuracy even further. In our tests, about 60% of fixations were in the calculated regions of salience (5% of the whole image), which corresponds to about 5 fixations within a time range of 2 seconds. We assume that there are three reasons which explain the improvement in

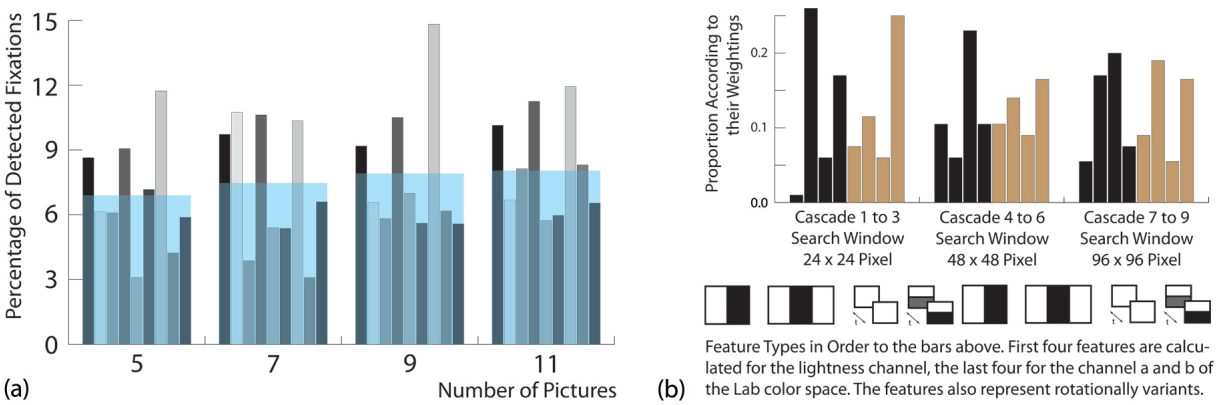

Fig. 7. (a) The graph shows how the 9 classifiers predict fixations trained with 5, 7, 9, and 11 pictures respectively. See Section 3 for how the detection of fixations was calculated. The transparent blue bars indicate the average values of the respective classifiers. (b) The graph shows the proportions of the chosen features by Adaboost taking into account the respective weightings.

prediction accuracy. Firstly, we run our tests with pictures of visual arts. Usually artists paint emotional contents in a distinctive and salient way, and therefore regions with high emotional salience coincide with visually salient regions. Secondly, as opposed to other approaches, our approach takes recent fixations into account and calculates a new salience map every 40 milliseconds. In this manner we are dealing with the real visual input of perception in each instance. Thirdly, to a limited extent, our method also implicitly trains object recognition (cf. 2.1). Therefore, the salience calculated by our attention model is often caused, for example, by people, faces, eyes, and hands, which are often associated with emotional salience.

4.2 Evaluation of the extended attention model

We used the recorded fixations of the second data set (Figure 6) to evaluate our extended attention model together with the simulation process. The data set includes the fixations of 14 participants viewing 10 pictures. For every picture, we created 14 simulations which we compared to the participants' fixations. The number of fixations per image and participant was between 40 and 110. For each simulation we took 100 fixations. If we define characteristic fixations as those which are fixated by at least 50% of the participants, the probability that these characteristic fixations are fixated on by a participant is at least 50%. According to our findings, these characteristic fixations also occur in our simulations with a frequency of at least 50%. Additionally, 95% of the characteristic fixations occurred in at least three simulations (out of a total of 14). We conclude that if we simulate enough fixations, we will get almost all characteristic fixations. The link between simulated and recorded fixations is shown by way of examples (Figure 8):

To compare the distribution of fixations of two scan paths, we calculate the sample correlation coefficient of the frequency values of the image points belonging to the fixations of the scan paths. While doing so, we project each fixation point of a scan path additive to a density map via a Gaussian window with a full-width at 96 pixels and a standard deviation of 32 pixels. We take the resulting distribution as the frequency of the fixations of a scan path. To compare two scan paths we normalize the distribution of fixations to an arithmetic mean of zero. The sample correlation coefficient derived from the set of the recorded fixation points serves as our measure for the similarity between two given fixation distributions.

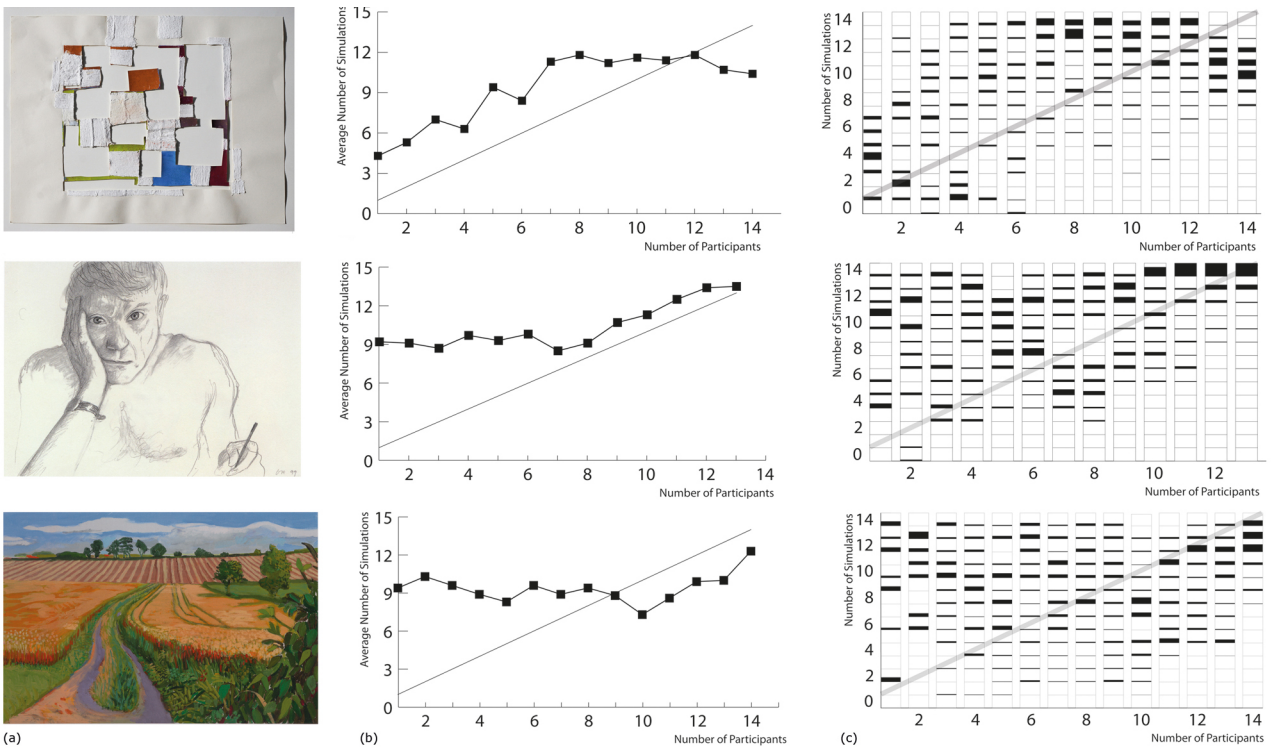

Fig. 8. (a) The picture viewed. (b) The horizontal axis refers to sets of fixations which are fixated on by a certain number of participants. The vertical axis shows the average number of simulations in which the fixations from one set occur. (c) The horizontal axis refers to sets of fixations which are fixated on by a specific number of participants. The vertical axis shows the number of simulations in which the fixations from one set occur, including their distribution. The black parts reflect the distribution. The diagonal in (b) and (c) facilitates orientation in the graph.

For each of the 10 pictures in data set II (Figure 6), we have 14 fixation sets recorded from the subjects and 14 sets of simulated fixations. For every fixation set, we calculated the frequency distribution for the corresponding picture and measured the similarity between the different fixation sets by calculating their sample correlation coefficient. The sample correlation coefficient between fixation sets recorded from subjects viewing the same picture was on average 0.55, and between recorded fixations and simulated fixations 0.43. Between simulated fixations the coefficient was 0.58; that is roughly the situation as if one person repeatedly looked at the same picture unaffected by the repetition, which is, of course, not possible in reality. The sample correlation coefficient between fixation sets recorded from subjects viewing different pictures was on average 0.3. The high correlation in this case was at first unexpected, but considering the fact that the most striking areas in most of the images are near the center, the middle vertical, and the mean horizontal, this high value is actually not surprising.

A comparison of the correlation between recorded and simulated fixations (0.43) and the correlation between recorded fixations (0.55) reveals that although a high correlation on the frequency between recorded and simulated fixations exists, the correlation between recorded data is significantly stronger. Whereas our simulation approximates the location of the fixations very closely, the approximation of the frequency is not as accurate. This can be an indication that frequency reflects the influence of cognition and individual factors more strongly than does the location of fixations, because cognition and individual factors cannot be simulated by our algorithm. This assumption is supported by the fact

that the correlation factors increase significantly if we do not compare the fixations of individuals but compare a total of 7 subjects each. In the latter case, the influence of individual factors diminishes. The comparison between recorded fixations resulted in a factor of 0.87, between recorded and simulated fixations in a factor of 0.55, and between recorded fixations and different pictures (with similar composition) in a factor of 0.44.

The procedure we used to investigate the dependence of the location of fixations on previous fixations is derived from a method presented by Privitera and Stark [2000]: Based on their location, the fixations of a scan path are grouped into 10 clusters. The 20 clusters of two scan paths are in turn grouped into 7 new clusters to compare these two scan paths. Now all fixations of the two scan paths can be assigned to one of the seven clusters, and the sequence of the fixations of the scan paths can be described by specifying the associated cluster numbers. Each of the seven clusters is assigned to a region in the picture. Therefore, the sequence of cluster numbers of a scan path describes the change of the regions during the observation of the image, and two scan paths can be declared as similar if the sequence of cluster numbers is similar. However, according to our observations, the sequence of fixations is characteristic of the behavior of a chaotic dynamic system. Therefore, the scan paths of two people looking at the same picture are very different after a few fixations, and there is no great benefit to investigating the entire sequences with respect to similarity. The distribution of fixations seems to be given for a certain person and image; however, there seem to be no canonical sequences of fixations. The distribution seems to be unaffected by occasional disturbances, e.g., if the subject's gaze leaves the image. In the simulation, the saliency of an image point depends on the current fixation, but the most striking points are like attractors in a chaotic system, so they are fixated on from time to time, and disturbances are automatically corrected. Therefore, the simulation shows behavior similar to that of subjects in this respect.

One can, however, examine how strongly the choice of a region/cluster depends on the previous region/cluster during the period of observation. This dependence can be described by a matrix whose elements state the quantity of changes from the cluster with the number of the line to the cluster with the number of the column. Calculated in this way, the sample correlation coefficient between the elements of the two matrices of two scan paths serves as a measure of similarity of the sequence of scan paths.

Our investigations lead to two results. First, the probability of a region being fixated on by a subject depends strongly on the preceding region. We calculated the frequencies of switching to a certain region in response to the previous region. The frequencies are given by the values of a column in the aforementioned matrix. The standard deviation as a percentage of the mean value of a target region is a measure of how strongly the choice of a region depends on the preceding region. The mean of this value of all regions in all pictures in data set II (Figure 6) was 134%. This high dependence on the predecessor shows the necessity to consider the respective preceding fixations by calculating a new fixation, just as the algorithm of Rajashekar et al. [2008] and our algorithm do. Second, the sample correlation coefficient between the matrices of subjects viewing the same picture was on average 0.52, and the coefficient between the matrices of subjects and the matrices of the simulated fixations was 0.38. This shows that the dependence between the simulated fixations just described was similar to the dependence between the recorded fixations, but the correlation between the recorded data was significantly stronger. If this difference was not due to individual or cognitive effects, there may be room for improving our extended attention model.

The example in Figure 9 describes the situation typical of many abstract paintings, as for example paintings by Christopher Wool or ink drawings by Henri Michaux. There are no meaningful regions or structures in these pictures, but people nevertheless recognize similarities or dissimilarities between them. Moreover, these pictures trigger certain associations and responses. For these pictures,

we wanted to know whether the distribution of fixations is governed by complex structures, in which case they could not be simulated realistically by our method; and whether the distribution of fixations possibly is a representation of those pictures' impact. First, everyone whom we showed the picture in Figure 9 immediately recognized the horizontal bar, along with the fact that this bar is the reason for disorder in the structure of the image. Still, even though the horizontal bar is the origin of the disorder in the image, this does not seem to affect the choice of fixations since the recorded fixations show no particular clustering of fixations at the interfering bar; nor did the interfering bar mark the beginning of the scan path exceptionally often. Second, all subjects recognized the disorder in the picture immediately, so they could not have had fixated on various bars at the moment they recognized the disorder. However, when they were given enough time, they fixated on all regions that are relevant to the disorder. It follows that the fixations, along with the pixels surrounding them, are suitable for a representation of the image's impact in this case.

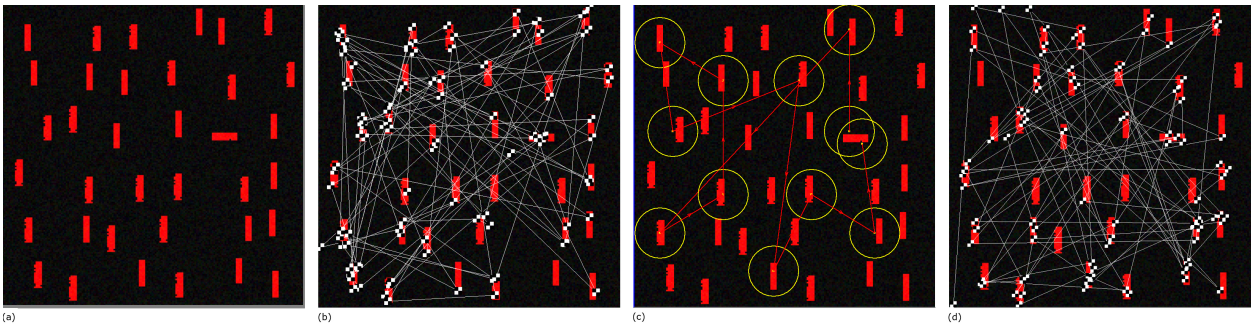

Fig. 9. (a) The image viewed. (b) Image with recorded fixations. (c) Prediction by Itti and Koch's algorithm. (d) Fixations simulated by our algorithm.

Since we often observed the effects just described with regard to abstract paintings, the example in Figure 9 illustrates the applicability of our algorithms - concerning both the simulation of fixations and the calculation of similarity - for abstract pictures. In addition, we observed that there was no preference for the center regions during the perception of abstract pictures, and that the scan paths followed no particular order, such as the observation from left-top to right-bottom.

The example discussed above was taken from the homepage of Itti and Koch [2013], where they present examples of simulated fixations due to calculated saliency maps. A comparison of their results with the recorded fixations shows that the salient regions detected by their algorithm are also regions with recorded fixations and can be taken as a representation of the picture. However, for our purposes, their algorithm does not yield a sufficient number of fixations. The order of the fixations that Itti and Koch's algorithm calculates does not so much describe the sequence of fixations; rather, it expresses the importance the fixation points have for perception. The recorded data shows that the order of initial fixations varied from one participant to the next, similar to the findings of our model.

We also evaluated our results by making the simulated fixations visible and assessing them based on experience. It is not yet known to what extent people can judge whether fixations are realistic. Therefore, human evaluation of simulated data is merely an indication of how realistic they are. However, there are applications where it is sufficient to assume that the fixations are taken to be realistic. Realistic avatar eye animation is such an application (cf. Itti and Dhavale [2003]). We tested our simulated fixations in a short experiment with nine people. Five of the participants had a lot of experience with eye tracker data and its analysis, while the other four had some experience with visual perception,

though not specifically with eye tracker data. The participants were presented 10 pictures each. For each picture they got two sets of 60 fixations. The fixations were visualized as shown in the following examples. The duration and the order of the fixations were given in a file. The participants had to decide which fixation sets were simulated, and which were recorded by an eye tracker. The mean value for the precision rate of the experienced participants was 72.8%. The inexperienced participants only reached a mean value of 53.5% and thus scored only little better than random decisions. As the examples in Figure 10 illustrate, it is not easy for people to decide which of the fixations featured in each instance are simulated and which have actually been recorded.

Fig. 10. Simulated and recorded fixations. Three pictures are shown twice. The difficulty in trying to figure out which of the fixations in the pictures above were simulated and which were recorded illustrates the consistency in the two sets of fixations.

4.3 Simulated fixations used for calculating similarity

We used simulated fixations for a prototype algorithm that calculates similarity. We assumed that fixations during the perception of visual arts along with the image information of their surroundings are important image information both for recognition and for comparing pictures. The points fixated on by a large number of people seemed particularly appropriate to us in this context. These “common fixations” are included in our simulated fixations as shown in Section 4.2. In order to determine the similarity between two given pictures, we calculated comparison features:

— The Euclidean distance of the mean coordinate values of the fixations, and the difference of the variance of the coordinates of the fixations.

— The difference of the average Euclidian distance and orientation of successive fixations, and the difference of the variance of the Euclidian distance and orientation of successive fixations.

— For every fixation, we calculated the mean value, the maximum contrast, granularity, and orientation of the dimension L of the L*a*b* color space within a radius of 32 pixels. The sample correlation

coefficients of these four variables, measured over all fixations of the pictures compared, were taken as comparison features. We did the same for the dimensions a and b of the L*a*b* color space.

We took the weighted sum of all comparison features as the value of similarity. We assumed that all features have an impact on how people perceive similarity, but we did not know how strong this influence is. Moreover, the range of the comparison features differed strongly. Therefore it was necessary to weigh the comparison features, but we did not have the selected criteria needed to choose the optimum weights. To deal with the issue of choosing appropriate weighting factors, we used two ways of calculating the comparison features that implicitly perform a weighting.

Approach 1: Based on our calculations, we want to organize a set of pictures according to the degree of similarity they have with a particular picture taken from the set, to be used as reference point. The first picture is the reference point, the second picture is the one most similar to it, and so on. We calculate all comparison features $x_{i,j}$ (i: feature number; j: picture id) between the selected picture and the rest of the set. For each of the 18 comparison features we calculate the mean value AM_i and the standard deviation σ_i. Then we calculate 18 new comparison factors $\tilde{x}_i$: $\tilde{x}_i = min((x_i - AM_i)/\sigma_i, 1)$. This approach takes the range of the similarity features into account, and limits the influence of any single feature in such a way that an individual alone cannot determine any great similarity.

Approach 2: As in Approach 1, we want to organize a set of pictures based on our calculations according to the degree of similarity between these pictures and a particular picture selected from the set. Unlike Approach 1, the given picture here is part of a subset of similar pictures. We calculate all comparison features among the pictures in the subset of similar pictures. Analogously to Approach 1, we calculate the mean value and the standard deviation for every comparison feature, and calculate 18 new comparison factors as in Approach 1. The assumption underlying this approach is that pictures that are similar to the pictures in the subset have similarity features which are within the range of the features in the subset. This leads to high values if the similarity features are within this range, and to low values if the similarity features are outside this range. Therefore, this approach factors in not only the range of similarity, as in Approach 1, but also similarity within the group of pictures in the given subset.

We used a corpus of about 2,500 pictures of visual arts for our study. One result using Approach 1 is that if a copy of a picture (with new simulated fixations) was inserted into the corpus, it was always identified as the most similar image. This was equally true when (at least 60) fixations derived from test subjects were used in the copy.

The following examples show exemplarily that the two approaches described above yield good results with regards to calculating similarity in terms of human perception.

Example 1: Our corpus features 30 photographs that depict the same person in different positions (Figure 11). There are about 500 portraits by Warhol in our corpus, about 150 of which are photographs of series similar to those in Figure 11, 54 with a similar color scheme. For each picture of the series, we sorted the pictures of our corpus by similarity using Approach 1. The 27 most similar pictures were always from the given series, and all 30 pictures were always among the 35 most similar pictures. If the pictures were randomly arranged, the probability that the first 27 pictures were pictures of the series was only about 10^{-52}.

Example 2: Our corpus contained 23 of Christine Gläser's abstract paintings. Selecting one of them (picture top left in Figure 12), we sorted the pictures of our corpus by similarity using Approach 1. Of the 23 pictures, 5 were among the 8 most similar. But not all pictures that our calculation showed to be similar were perceived as being similar by human viewers. To improve the results, we used Approach 2. For a given picture, the user first sorts the pictures of our corpus by similarity, using Approach 1. Next, he selects from among the most similar pictures, the pictures he perceived as being most similar.

Fig. 11. A subset of the pictures used in Example 1.

From the latter, he chooses one picture and then sorts the corpus again using Algorithm 2. This process can be executed several times in succession until the result no longer improves. Figure 12 shows the results of the first iteration of this method. After this step, the 8 most similar pictures were all pictures of Gläser's abstract paintings.

Fig. 12. Top row: The picture on the far left is the point of reference compared to which the pictures of our corpus were sorted by similarity (Approach 1), followed by the 8 most similar pictures. Bottom row: The picture on the far left is the point of reference compared to which the pictures of our corpus were sorted by similarity, now using Approach 2; followed by the 5 pictures selected by the user as being similar (marked with white squares); and the most similar pictures.

5. CONCLUSIONS

Everyday situations show that it takes people only a short period of time to recognize pictures. In the case of pictures of visual arts, we usually need less than two seconds to decide whether or not we know a given picture. If, for example, people have some experience with pictures by Picasso and Van Gogh, they can assign pictures to these artists even before they realize the content in detail, and they can assign pictures to these artists even if they have not seen the pictures in question before.

Recognition and perceiving similarity occur in such a short time span that there is no space for elaborate investigation. We assume that there are basic visual features in pictures that determine their appearance but are not necessarily linked to what they express. We claim that these basic visual features, in contrast to the more complex picture contents, are accessible to automatic analysis and may be used for analyzing large volumes of pictures. Thus, for example, algorithms may explore potential links, tendencies, and developments over time.

As our findings show, 70% of the fixations of one subject are also fixated on by more than 50% of all subjects, and we assume that these fixations describe the desired features. We call these fixations "common fixations"; according to our results, these features are also among our simulated fixations.

Based on the simulated fixations of every picture in our art corpus, we developed a straightforward algorithm to calculate similarity among pictures. If a copy of a picture (with new simulated fixations) was inserted into the corpus, it was always identified as the most similar image, which is the best

obtainable result for the purpose of identifying pictures. The evaluation of our similarity calculations revealed that our algorithm finds most of the pictures that are similar to a given picture according to human perception; nevertheless, not every picture that our calculations showed to be similar to the given picture was perceived as being similar by human viewers. However, interaction enables users to easily find groups of similar pictures. The results of our similarity calculations confirm, that in fact especially common fixations characterize pictures of visual arts to a large extent.

As already mentioned, the common fixations of subjects are also included among the simulated fixations. On the other hand, the correlation of the frequency of fixations among the subjects is stronger than between fixations of subjects and simulated fixations. However, the simulated fixations are realistic enough and need not be further improved for our similarity calculations. Our findings revealed that if simulated fixations are substituted with recorded fixations, there are no significant differences in the results. The simulated fixations are also sufficient to simulate eye movements (such as an avatar's), because only experts can distinguish better than chance between the simulated and the recorded data. In this case, our algorithm can also be applied to videos, an undertaking which would, however, require that the classifiers be trained beforehand with the appropriate date.

It is uncertain whether the simulation of fixations can be further improved, since the location and the frequency of fixations are influenced by cognition and individual factors, and their influence can hence not be directly predicted based on visual input. However, it is plausible that specialized training data produce better results in restricted picture sets. It seems possible that special classifiers could be trained, for instance for portraits, landscapes, graphics, or images outside the arts. It is also plausible to train classifiers for special tasks, for example, to predict the time a subject needs to fixate on all important regions in a diagram.

Our goal is to analyze large numbers of pictures and to make them explorable. For this purpose, we want to automatically examine picture quantities for potential links, tendencies, and developments over time. We are also interested in realizing structures and ordering principles in terms of gestalt psychology. We assume that the calculation of realistic fixations described in this article is a useful tool for detecting basic features in pictures that can be used for these objectives.

Both the simulation of fixations and the algorithm for the calculation of similarity have a linear complexity. We use a PC with an Intel i7-2600 processor with 3.4 GHz and 16 GB computer memory. With this computer equipment, the simulation of fixations runs in real time, and the algorithm for the calculation of similarity requires two seconds to rank our art corpus with 2,500 pictures according to similarity. This performance is quite sufficient for our purposes, so that we felt we did not need any further optimization. Using our algorithm for calculating similarity and displaying the results on thumbnails only requires the memory space of 40 KByte per picture; hence memory space does not constitute an issue for our algorithms. However, both algorithms can be optimized so that one could interactively work with them in an art corpus with millions of pictures, such as for example the "Bildarchiv Prometheus". We could also parallelize the algorithms to a high degree; moreover, they are suited to run on graphic processors. Therefore, they could also be optimized for even larger corpora.

REFERENCES

2014. Feng Shui for Graphic User Interfaces. Internet. (2014). http://www.feng-gui.com/ Retrieved 2014.08.18.

Mohamed Alaa El-Dien Mahmoud Hussein Aly. 2011. *Searching Large-Scale Image Collections*. Ph.D. Dissertation. California Institute of Technology.

James W. Davis, Alexander M. Morison, and David D. Woods. 2007. An adaptive focus-of-attention model for video surveillance and monitoring. *Machine Vision and Applications* 18, 1 (2007), 41–64.

Benjamin Höferlin, Hermann Pflüger, Markus Höferlin, Gunther Heidemann, and Daniel Weiskopf. 2012. Learning a Visual Attention Model for Adaptive Fast-forward in Video Surveillance. In *ICPRAM (2)'12*. 25–32.

James Hoffman and Baskaran Subramaniam. 1995. The role of visual attention in saccadic eye movements. *Perception & Psychophysics* 57 (1995), 6: 787–795.

Kenneth Holmqvist, Marcus Nyström, Richard Andersson, Richard Dewhurst, Halszka Jarodzka, and Joost van de Weijer. 2010. *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.

Laurent Itti. 2005. Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition* 12 (2005), 1093–1123.

Laurent Itti and Nitin Dhavale. 2003. Realistic avatar eye and head animation using a neurobiological model of visual attention. In *Proceedings of SPIE*. SPIE Press, 64–78.

Laurent Itti and Christof Koch. 2013. Bottom-Up Visual Attention Home Page; The Interactive Demo. Internet. (2013). http://ilab.usc.edu/bu/javaDemo/ie/demo.html Retrieved 2013.08.18.

Laurent Itti, Christof Koch, and Ernst Niebur. 1998. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* 20, 11 (1998), 1254–1259.

Hector Jasso and Jochen Triesch. 2007. Learning to Attend - From Bottom-Up to Top-Down. In *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint*, Lucas Paletta and Erich Rome (Eds.). Lecture Notes in Computer Science, Vol. 4840. Springer Berlin Heidelberg, 106–122.

Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. 2009. Learning to predict where humans look.. In *Computer Vision, 2009 IEEE 12th International Conference on Computer Vision*. IEEE, 2106–2113.

Wolf Kienzle, Bernhard Schölkopf, Felix A. Wichmann, and Matthias O. Franz. 2007. How to find interesting locations in video: a spatiotemporal interest point detector learned from human eye movements. In *Proceedings of the 29th DAGM conference on Pattern recognition*. Springer-Verlag, Berlin, Heidelberg, 405–414.

Krystian Mikolajczyk, Tinne Tuytelaars, Cordelia Schmid, Andrew Zisserman, Jiri Matas, Frederik Schaffalitzky, Timor Kadir, and Luc van Gool. 2005. A Comparison of Affine Region Detectors. *International Journal of Computer Vision* 65 (2005), 1/2 43–72.

Frieder Nake. 1974. *Ästhetik als Informationsverarbeitung*. Springer Verlag Wien New York.

Sunaad Nataraju, Vineeth Balasubramanian, and Sethuraman Panchanathan. 2009. Learning attention based saliency in videos from human eye movements. In *Proceedings of the 2009 international conference on Motion and video computing (WMVC'09)*. IEEE Computer Society, Washington, DC, USA, 134–139.

Yaqing Niu, Rebecca M. Todd, Matthew Kyan, and Adam K. Anderson. 2012. Visual and emotional salience influence eye movements. *ACM Trans. Appl. Percept.* 9, 3, Article 13 (2012), 18 pages.

Robert J. Peters and Laurent Itti. 2007. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*. 1–8.

Claudio M. Privitera and Lawrence W. Stark. 2000. Algorithms for Defining Visual Regions-of-Interest: Comparison with Eye Fixations. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 9 (2000), 970 – 982.

Umesh Rajashekar, Ian van der Linde, Alan C. Bovik, and Lawrence K. Cormack. 2008. GAFFE: A Gaze-Attentive Fixation Finding Engine. *IEEE Transactions on Image Processing* 17 (2008), 564 – 573.

Robert J. Peters. W. Rodieck. 1998. *The First Steps in Seeing*. Sinauer Associates, Inc.

Jukka Saarinen. 1993. Shifts of Visual Attention at Fixation and Away from Fixation. *Vision Research* 33 (1993), 8: 1113–1117.

Paul Viola and Michael J. Jones. 2004. Robust Real-Time Face Detection. *Int. J. Comput. Vision* 57, 2 (2004), 137–154.

Jeremy M. Wolfe. 1994. Guided Search 2.0 A revised model of visual search. *Psychonomic Bulletin & Review* 1, 2 (1994), 202–238.

Qi Zhao and C. Koch. 2011. Learning visual saliency. In *Information Sciences and Systems (CISS), 2011 45th Annual Conference on*. 1–6.

PHOTOGRAPHS

David Hockney, "Path Through Wheat Field, July," 2005 (Figures 6, 8, and 10) and "The Sermon on the Mount III (after Claude)," 2010 (Figure 6) ©David Hockney, photo credit: Richard Schmidt.
David Hockney, "Self-Portrait, Baden-Baden, 8th June 1999" (Figures 6, 8, and 10) ©David Hockney, photo credit: Steve Oliver Collection The David Hockney Foundation.
All other photographs were created by Hermann Pflüger with permission of the artists or owners, respectively.